Fluctuations, irreversibility and causal influence in time series.

DISSERTATION
zur Erlangung des akademischen Grades
Doctor of Philosophy
(Ph.D.)

im Fach
Biophysik

eingereicht an der
Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
M.Sc. Andrea Auconi

Präsidentin/Präsident der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekanin/Dekan der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin:
Prof. Dr. Bernhard Grimm

Gutachter/innen:

1. Prof. Dr. Dr. h.c. Edda Klipp
2. Prof. Dr. Igor Sokolov
3. Prof. Dr. Benjamin Lindner

Tag der mündlichen Prüfung: 02.04.2019

https://doi.org/10.18452/19949

Abstract

Information thermodynamics is the current trend in statistical physics. It is the theoretical research of a unified framework for the description of nonequilibrium features of stochastic dynamical systems like work dissipation and the irreversibility of trajectories, using the language of fluctuation theorems and information theory.

The model-independent nature of information and irreversibility allows a wide applicability of the theory to more general (nonphysical) models from systems biology and quantitative finance, where asymmetric interactions and nonlinearities are ubiquitous. In particular, we are interested in time series obtained from measurements or resulting from a time discretization of continuous models.

In this thesis we study the irreversibility of time series considering the statistical properties of their time-reversal, and we derive a fluctuation theorem that holds for time series of signal-response models, and that links irreversibility and conditional information towards past.

Interacting systems continuously share information while influencing each other dynamics. Intuitively, the causal influence is the effect of those interactions observed in terms of information flow over time, but its quantitative definition is still under debate in the community. In particular, we adopt the scheme of partial information decomposition (PID), that was recently defined in the attempt to remove synergistic and redundant effects from information-theoretic measures. Here we propose our PID, and motivate the resulting definition of causal influence for the special case of linear signal-response models.

The thermodynamic role of causal influences can only be discussed for time series of linear signal-response models in the continuous limit, and its generalization to general time series remains in our opinion the open problem in information thermodynamics.

Finally, we apply the time series irreversibility framework to quantify the dynamical response to perturbations in a molecular model of circadian rhythms. It results that, while the dynamics is mostly oscillating with 24h period, the response to perturbations measured by a mutual mapping irreversibility measure is highly characterized by 12h harmonics.

Zusammenfassung

Informationsthermodynamik ist der aktuelle Trend in der statistischen Physik. Es ist die theoretische Konstruktion eines einheitlichen Rahmens für die Beschreibung der Nichtgleichgewichtsmerkmale stochastischer dynamischer Systeme, wie die Dissipation der Arbeit und die Irreversibilität von Trajektorien, unter Verwendung der Sprache der Fluktuationstheoreme und der Informationstheorie.

Die modellunabhängige Natur von Information und Irreversibilität ermöglicht eine breite Anwendbarkeit der Theorie auf allgemeinere (nichtphysikalische) Modelle aus der Systembiologie und der quantitativen Finanzmathematik, in denen asymmetrische Wechselwirkungen und Nichtlinearitäten allgegenwärtig sind. Insbesondere interessieren wir uns für Zeitreihe, die aus Messungen gewonnen werden oder aus einer Zeitdiskretisierung kontinuierlicher Modelle resultieren.

In dieser Arbeit untersuchen wir die Irreversibilität von Zeitreihen unter Berücksichtigung der statistischen Eigenschaften ihrer Zeitumkehrung, und leiten daraus ein Fluktuationstheorem ab, das für Signal-Antwort-Modelle gilt, und das Irreversibilität sowie bedingte Informationen mit der Vergangenheit verknüpft.

Interagierende Systeme tauschen kontinuierlich Informationen aus und beeinflussen sich gegenseitig. Intuitiv ist der kausale Einfluss der Effekt dieser Wechselwirkungen, der im Hinblick auf den Informationsfluss über die Zeit beobachtet werden kann, aber seine quantitative Definition wird in der Fachgemeinschaft immer noch diskutiert. Wir wenden insbesondere das Schema der partiellen Informationszerlegung (PID) an, das kürzlich definiert wurde, um synergistische und redundante Effekte aus informationstheoretischen Maßen zu entfernen. Hier schlagen wir unsere PID vor und diskutieren die resultierende Definition des kausalen Einflusses für den Sonderfall linearer Signal-Antwort-Modelle.

Die thermodynamische Rolle kausaler Einflüsse kann nur für lineare Signal-Antwort-Modelle in der kontinuierlichen Grenze diskutiert werden, und ihre Verallgemeinerung auf allgemeine Zeitverläufe bleibt nach unserem Erachten das offene Problem in der Informationsthermodynamik.

Schließlich wenden wir Informationsthermodynamik von Zeitreihen an, um die dynamische Reaktion auf Störungen in einem molekularen Modell zirkadianer Rhythmen zu quantifizieren. Dies führt dazu, dass, während die Dynamik meistens im 24-Stunden-Takt oszilliert, die Reaktion auf Störungen stark von 12-Stunden-Rhythmen bestimmt wird.

> *Zarathustra entgegnete: "Was erschrickst du desshalb? - Aber es ist mit dem Menschen wie mit dem Baume. Je mehr er hinauf in die Hoehe und Helle will, um so staerker streben seine Wurzeln erdwaerts, abwaerts, in's Dunkle, Tiefe, - in's Boese."*
>
> **— Friedrich Nietzsche**

(EN) Zarathustra answered: "Why art thou frightened on that account?—But it is the same with man as with the tree. The more he seeketh to rise into the height and light, the more vigorously do his roots struggle earthward, downward, into the dark and deep—into the evil." [Ger12; Nie98]

Contents

1 Introduction

1.1 Motivation and problem statement

Data from measurements on dynamical systems are given in the form of time series[Ham94]. Examples are the electrical activity of the heart[Cos+02; Gol+00], seismological data describing earthquakes[Oga88], real-time gene expression[Giu+01], exchange rates in the currency market[Tay08], and many more. These are all "complex" systems, meaning that are observed at a macroscopic level, and the lack of information on the microscopic details results in an uncertainty whose effect is non negligible. Therefore probability is introduced in descriptions, and in particular stochastic processes are used to generate probabilistic dynamics.

Physical observations (measurements) on real systems have necessarily a finite time-resolution, and continuous trajectories are just build up as a mathematical abstraction. We will anyway assume an underlying time-continuous dynamics without loss in generality, and we will consider time series as a result of observing trajectories at a finite frequency $\frac{1}{\tau}$. The *observational time* τ can be considered as the fundamental parameter in this thesis, because it specifies the time series framework. While in experimental data the observational time τ has a lower bound given by the intrinsic sampling rate of the measuring instrument and can only be increased, in stochastic dynamical models it can be arbitrarily varied and it gives insights on the model system behavior at different time scales.

We are interested in interacting systems, and we wish to characterize and quantify the effect of interactions in terms of *information flow*. While there is quite general agreement in the definition of the mutual information between variables in terms of entropy measures[CT12], the definition of information flow is still debated[Jam+16]. The mutual information[CT12] was initially used as a measure of information flow in the characterization of small signaling networks[Tka+09; Wal+10; Tka+12a; ST15] and neural codes[Bia17; Str+98; DB95] in theoretical neuroscience[Day+01]. The mutual information is a symmetric measure, while information flow has to be associated with a directionality. Therefore directed information and conditional mutual information (or transfer entropy) measures were introduced[Mas90; Sch00].

Transfer entropy is based on conditioning and is dominated by synergistic effects, therefore it was criticized as a measure of information flow[Jam+16]. Consequently the framework of partial information decomposition (PID) was introduced as an attempt to define a measure of "unique" non-redundant and non-synergistic information flow[Rau+14; WB10], that is what we call the *causal influence* [Auc+17]. Many schemes of PID were proposed, but they were all shown to have unsatisfactory features already for Gaussian systems[Bar15]. In particular, they produce threshold effects that do not seem an appropriate description of linear dynamics.

The quantitative definition of a causal influence measure is motivated by a problem of scientific communication of results. Really often debates arise on the statistical significance of claims about causal influences between observables. This often results in superficial critics of correlations and Granger causality measures in data analysis, often concluded with the imprecise and pretentious statement: "correlation is not causation". Those kind of discussions will never result in a well-posed problem until a quantitative definition of causal influence is provided. If the community will agree on a particular definition, then the discussion about causal influence in finite (maybe small) time series data will be mainly focused on statistics.

Mutual information and transfer entropy are fundamental quantities in modern thermodynamics, because they provide bounds on dissipation and extracted work in feedback systems[Sei12; SU12; IS13]. Indeed, the information thermodynamics community has recently focused (again) on *fluctuations*, meaning the statistics of single realizations, and also the fluctuations of information measures have been considered. This led to the development of integral fluctuation theorems (IFTs) linking entropy production and information measures in stochastic systems[Par+15; RH16; IS13; Ito16]. Note that in nonequilibrium steady-states of Markovian systems the entropy production is directly linked to the macroscopic dissipation[Sei12], or to the extracted work in controlled systems[RH16]. In general, these IFTs are all based on continuous stochastic descriptions (or on Hamiltonian descriptions with initial states sampled from canonical distributions[Sag11]) that allow the identification and additive separation of the heat exchanged with thermal baths by different subsystems. This corresponds to the bipartite (or multipartite) structure, where the evolution of different variables are independent in updating. For a bidimensional system this is written $p(x_{t+dt}, y_{t+dt}|x_t, y_t) = p(x_{t+dt}|x_t, y_t) \cdot p(y_{t+dt}|x_t, y_t)$.

The key observation is that time series are not bipartite. Indeed for a finite observational time $\tau > 0$ it holds $p(x_{t+\tau}, y_{t+\tau}|x_t, y_t) = p(x_{t+\tau}|x_t, y_t) \cdot p(y_{t+\tau}|x_t, y_t, x_{t+\tau})$, meaning that the anti-causal effect of $x_{t+\tau}$ in the prediction of $y_{t+\tau}$ is not negligible. Therefore information thermodynamics theory requires a different formalization for time series.

Irreversibility is defined as the statistical distinguishability between a process and its time reversal conjugate[RP12], it is related to dissipation in thermodynamics[Kaw+07], and is in general a symptom of nonlinear dynamics[Por+07]. While the quantification of the irreversibility in time series is an already active field of research[RP12; Lac+12], its connection to information-theoretic measures and fluctuations was not considered so far. Such study will likely be helpful for the general problem of modeling dynamical systems from time series data, and ultimately for their prediction and control[Sti+12].

We will be mainly interested in signal-response models, defined as stationary stochastic processes characterized by the absence of feedback.

The three major achievements of this PhD thesis are stated here:
(1) We developed an information thermodynamics framework for bivariate time series, and identified an integral fluctuation theorem for signal-response models.
(2) We defined a quantitative measure of causal influence for linear signal-response models.
(3) We proposed a quantitative characterization of the influence of photic perturbations on circadian rhythms, that is based on ideas from points (1) and (2).

The causal influence measure is published in [Auc+17] and discussed in Chapter 4. The fluctuation theorem for time series is published in [Auc+19b], and discussed in Chapter 5. Preliminary results of the information thermodynamics framework applied to a model of circadian oscillations is discussed in Chapter 6, and submitted for publication in Phys Rev E (Preview available in [Auc+19a]).

The aim of future research would be to generalize the causal influence measure to general systems with feedbacks and nonlinearities, and to clarify its role in the information thermodynamics of time series. This would amount to a general relation between causality and the irreversibility of time series.

Finally we note that, even if we took a different direction with the study of time series, the information thermodynamics application to multipartite Bayesian networks developed by Sosuke Ito[IS13] was a great inspiration to this thesis.

1.2 Thesis Structure and Summary

1.2.1 Chapter 2

Stochastic differential equations

Stochastic differential equations (SDEs) are the basic ingredient for the models considered in this thesis. Here we introduce the fundamental properties of Brownian motion and of the Ornstein-Uhlenbeck process, that are the simplest descriptions of fluctuations. They will be used in the following chapters as input signals or perturbations to the more structured models that will be considered. Then we will introduce the spectral analysis of stochastic processes, that is particularly useful in the description of oscillations, and we will take the stochastic linear negative feedback loop as the basic example. We will derive the equivalent description of SDEs in terms of the Fokker-Planck partial differential equation. The two different interpretations of SDEs by Ito and Stratonovich and the resulting differences in stochastic calculus and path integrals are discussed.

1.2.2 Chapter 3

Information thermodynamics on bipartite systems

We review the modern theory of information thermodynamics as it was developed in the last 6-8 years, with critical discussions. We start from the basic definitions of mutual information and transfer entropy, and their introduction in relevant biological problems[Tka+08b; IS15]. Then we introduce the thermodynamics quantities in nonequilibrium statistical mechanics discussing the Jarzynski-Crooks fluctuation theorems[Jar11; Kur98; Cro99], and the generalization to measurement and feedback systems[SU12; SU09]. Then we introduce the study of fluctuations discussing the two main recent approaches: the Horowitz-Esposito information flow[HE14], and the Ito-Sagawa transfer entropy inequalities[Ito16; IS13]. The theory is considered quite in detail, and it is the basis to understand the novelty of our contribution to the field of information thermodynamics, that is the extension of integral fluctuation theorems to (non-bipartite) time series, and is discussed in Chapter 5.

Importantly, we note that very recently a new manuscript[Ito18] appeared in the ArXiv providing a unified framework for the second law of information thermodynamics, that is based on information geometry[Ama97] and incorporates most of the results obtained for bipartite systems discussed in this chapter.

1.2.3 Chapter 4

Causal influence

We introduce linear signal-response models, and study the information processing properties of the basic linear response model (BLRM) providing analytical results. We adopt the partial information decomposition framework[Bar15], we define our measure of redundancy, and this leads to our definition of causal influence $C_{x\to y}(\tau)$.

We discuss the τ dependence of $C_{x\to y}(\tau)$ and argue that it has the intuitive properties of a causal influence measure: it is a peak function starting from $C_{x\to y}(0) = 0$ and vanishing for $\lim_{\tau\to 0} C_{x\to y}(\tau) = 0$ meaning that the effect of interactions is observed gradually over time and then disappears after long enough time. It is zero in the absence of direct (or mediated) interaction, and can be generalized to multidimensional systems. Difficulties arising in the generalization to feedback systems are discussed.

1.2.4 Chapter 5

Information thermodynamics on time series

We set the basis for the study of fluctuations, information flow, and irreversibility in bivariate time series. Time series result from a time discretization with sampling interval τ of a continuous underlying dynamics. We define causal representations of time series obtained from stochastic dynamics, highlighting their non-bipartite structure.

We define the measure of "mapping irreversibility" to describe the statistical properties of transitions over single τ intervals, and it is a Markov approximation to the irreversibility of a whole time series[RP12]. In signal-response models the "mapping irreversibility" is equivalent to the irreversibility of a whole time series.

In the case of (nonlinear) signal-response models we found a fluctuation theorem, and a corresponding inequality that sets the backward transfer entropy[Ito16] as a lower bound to the conditional mapping irreversibility. We verified such inequality in the BLRM obtaining analytical solutions, and in a nonlinear biological model of receptor-ligand systems. There we showed the importance of fine-tuning the observational time. We also introduced a measure of irreversibility density, that describes the microscopic configurations that contribute more to the irreversibility of the macroscopic process.

In the BLRM the backward transfer entropy converges to the Horowitz-Esposito information flow[HE14] and to our causal influence rate in the limit $\tau \to 0$, this being a first hint on the role of causal influences in thermodynamics.

1.2.5 Chapter 6

Quantifying the effect of photic perturbations on circadian rhythms

We consider a deterministic delay differential equation model of the circadian core clock network introduced in [Kor+14]. That is a minimal model of circadian oscillations and it reproduces amplitudes and phase shifts between genes observed in mammalian tissues. We know from literature how the circadian clock responds to pulse-like perturbations in terms of phase-response curves[Gra+09], and also at a molecular level as an activation of *Period* genes[GR10; Yan09; RW01].

We investigated continuous photic perturbations leading the system out of equilibrium, and studied the irreversibility of the resulting time series. We propose to characterize the light-induced response with a measure of mutual irreversibility, defined as the Markovian approximation to the mutual entropy production introduced in [DE14].

While the circadian clock dynamics is basically 24h oscillations, the response to light perturbations measured by the mutual mapping irreversibility results to be strongly characterized by 12h harmonics.

2 Stochastic differential equations

2.1 Why probabilistic descriptions?

Probability is the language to describe incomplete knowledge in the prediction of future events.

A popular example is the coin tossing experiment, head or tails, for which one assumes a $\frac{1}{2}$ probability for both possible outcomes. Nevertheless, in a classical mechanics experiment like this, if we have sufficiently precise information on the state of the coin immediately after it is launched (speed and angular momentum), then it is just a matter of calculus to predict the outcome of the experiment with almost no uncertainty. Of course this is never the case that we wish to predict the outcome of the coin toss, while we just accept the uncertainty of the outcome that is necessarily there if no information on the initial state of the coin is considered. The experiment is then called Random, unpredictable.

In general, even if we agree on the theoretical possibility of a deterministic description of nature, all the mathematical models for the dynamics of physical observables are appropriate descriptions only for some spatial and time scales that are close to the observer's perspective, and they will never be complete descriptions of nature. Furthermore, even if we assume that a mathematical model is the complete description of a physical phenomenon, still there can be sensitive dependence on initial conditions, and forecasting on longer time scales may require an increasing level of detail on the specification of the initial state of the variables. This leads to an effective unpredictability which is widely discussed in chaos theory[Vul10] starting from the Lorenz model for atmospheric convection[Lor63].

We deal with uncertainty on predictions in many real world situations: sport competitions, weather forecasts, political elections outcomes, projections of population growth in a city... just to name a few. In all these cases, probabilistic descriptions are used to express the information one has on those systems in terms of the likelihood of particular future events to occur. Probability is a so intuitive and natural concept in human language that some scientist even suggested that the world could be inherently probabilistic[Dir81]. This is always the case for quantum mechanics,

where the Heisenberg uncertainty principle forces the description of systems states to be probabilistic.

Whatever the origins of the uncertainty on predictions is, probabilistic descriptions are widely used in physics and engeneering since the introduction of statistical ensembles in thermodynamics by Gibbs[Gib14; Hua87] and the study of Brownian motion by Einstein[Ein56]. In biophysics, input-output probabilistic models were the natural choice for the study of information processing in transcriptional regulation and signaling pathways[Tka+08b; Kli+16; Tka+09; Wal+10; Tka+12b]. Random trajectories and fluctuations described by stochastic differential equations are used in quantitative finance for the asset dynamics and the corresponding derivative pricing[Shr12; Shr04].

Here we introduce the formalism of continuous-time stochastic processes, that will be our preliminary framework for the study of applications like the receptor-ligand system and the circadian clock network model. We first define Brownian Motion and the Ornstein-Uhlenbeck process, which are the most popular descriptions of fluctuations, and then we consider more in general stochastic differential equations (SDEs) in the Ito and Stratonovich interpretations. Only from chapter 4 we will discuss the concept of observational time and the construction of time series from SDEs.

2.2 Brownian Motion

Brownian motion $W(t)$ is the fundamental continuous-time stochastic process. It is defined[Shr04] by the following properties:

1) $W(t)$ *is almost surely a continuous function of time.*
2) $W(0) = 0$.
3) $W(t)$ *has independent increments, meaning that for any partition* $0 = t_0 < t_1 < \ ... \ < t_{m-1} < t_m$ *the increments* $W(t_1), W(t_2) - W(t_1), \ ... \ , W(t_m) - W(t_{m-1})$ *are independent.*
4) for any partition $0 = t_0 < t_1 < \ ... \ < t_{m-1} < t_m$ *each increment is normally distributed with* $\langle W(t_i) - W(t_{i-1}) \rangle = 0$ *and* $\langle (W(t_i) - W(t_{i-1}))^2 \rangle = t_i - t_{i-1}$.

The brakets $\langle \ \rangle$ indicate ensemble averages. Here the term "*almost surely*" indicate that a statement is true with probability $P = 1$, it is satisfied for all possible realizations of the process except for a set of realizations with vanishing probability.

It is clear that for any time interval $T > 0$ there are infinitely many possible realizations of the Brownian Motion $W(t)$ with $0 \leq t \leq T$ and the amount of information needed to describe one of these stochastic trajectories diverges with the precision of the specification. We will always be interested in ensemble properties of stochastic processes rather than in single realizations. Nevertheless, there is one property of Brownian Motion that is true both at the ensemble and single trajectory level, and it is its ***nonzero quadratic variation***.

The quadratic variation up to time T of a function $f(t)$ is defined as:

$$[f, f](T) \equiv \lim_{||\Pi|| \to 0} \sum_{j=0}^{n-1} [f(t_{j+1}) - f(t_j)]^2, \tag{2.1}$$

where $0 = t_0 < t_1 < ... < t_{n-1} < t_n = T$, and $||\Pi|| = \max_i [t_{i+1} - t_i]$.

If the function f has a continuous derivative then there exist a point t_j^* in the subinterval $[t_j, t_{j+1}]$ such that $f(t_{j+1}) - f(t_j) = f'(t_j^*) \cdot (t_{j+1} - t_j)$. Then for the quadratic variation we have:

$$\begin{aligned} [f, f](T) &\leq \lim_{||\Pi|| \to 0} \left[||\Pi|| \cdot \sum_{j=0}^{n-1} |f'(t_j^*)|^2 (t_{j+1} - t_j) \right] = \\ &= \lim_{||\Pi|| \to 0} ||\Pi|| \cdot \int_0^T |f'(t)|^2 dt = 0, \end{aligned} \tag{2.2}$$

where in the last passage we use the fact that the integral $\int_0^T |f'(t)|^2 dt$ is finite since $f'(t)$ is continuous on $[0, T]$ and therefore bounded in this interval.

All the functions with continuous derivative have zero quadratic variation. This is not the case for Brownian Motion, because the ratio $\frac{W(t_{j+1}) - W(t_j)}{t_{j+1} - t_j}$ almost surely diverges in the limit $t_{j+1} - t_j \to 0$ and the derivative W' is not defined. The Brownian Motion velocity can still be defined as uncorrelated white noise using the Dirac delta distribution as it is often the case in the physics literature, but the white noise is not continuous because of the independent increments property, and then the argument in equation (2.2) does not hold. The Brownian Motion has the important property that it accumulates quadratic variation at rate one per unit time, meaning that the value calculated on a single realization of the dynamics is equal to $[f, f](T) = T$ almost surely, i.e. it is for all possible realizations of Brownian Motion $W(t)$ except for a set of realizations with zero probability.

To show this we consider a sampled quadratic variation Q_Π corresponding to a particular partition Π:

$$Q_\Pi \equiv \sum_{j=0}^{n-1} \left(W(t_{j+1}) - W(t_j) \right)^2 . \tag{2.3}$$

We have to show that in the limit $||\Pi|| \to 0$ the expected value of Q_Π is T and its variance converges to zero. The expected value is $\langle Q_\Pi \rangle = T$ for every partition Π as it is directly implied by the second moment of the increments distribution in the definition of Brownian Motion. Let us now recall that for a normal Random variable N with zero mean it holds $\langle N^4 \rangle = 3(\langle N^2 \rangle)^2$. Then for the variance of Q_Π we have:

$$\begin{aligned}\langle (Q_\Pi - \langle Q_\Pi \rangle)^2 \rangle = \sum_{j=0}^{n-1} \left\langle \Big((W(t_{j+1}) - W(t_j))^2 - (t_{j+1} - t_j) \Big)^2 \right\rangle = \\ = \sum_{j=0}^{n-1} 2(t_{j+1} - t_j)^2 \leq 2||\Pi|| T,\end{aligned} \tag{2.4}$$

which converges to 0 as $||\Pi|| \to 0$. Then we have shown that the quadratic variation of Brownian motion is:

$$[W, W](T) = T. \tag{2.5}$$

2.2.1 The white noise representation

Let us define white noise Γ as the limit for $dt \to 0$ of a normal Random variable with mean 0 and variance $\frac{1}{dt}$, $N(0, \frac{1}{dt})$. This is is related to the definition of Brownian motion through $dW = \Gamma \, dt$. The white noise Γ is not continuous because it is proportional to the increments of Brownian motion which are independent. As a consequence of the nonzero quadratic variation of Brownian motion (Eq.2.5), the white noise Γ diverges almost surely in the limit $dt \to 0$ and, like the Dirac delta, it belongs to the class of distributions and has a meaning only under the integral sign. We can write Brownian motion as a function of the particular realization of Γ:

$$W(t) = \int_0^t \Gamma(t') \, dt'. \tag{2.6}$$

The properties of Γ follow directly from the discussion of the Brownian motion increments:

$$\langle \Gamma(t) \rangle = 0 \quad \forall t \tag{2.7}$$

$$\langle \Gamma(t) \Gamma(t + t') \rangle = \delta(t'), \tag{2.8}$$

where $\delta(t')$ is the Dirac delta distribution.

This representation is the most commonly used in physics, especially in the path integral framework[Sei12]. The use of this formalism implicitly chooses the Stratonovich interpretation of stochastic differential equations, while expressions in terms of Brownian motion dW correspond to the Ito interpretation. We will discuss the Ito-Stratonovch difference in the description of multiplicative noise, while for the

description of Brownian motion and of the Ornstein-Uhlenbeck process the two interpretations are equivalent.

2.3 Ornstein-Uhlenbeck process

The Ornstein-Uhlenbeck process $x(t)$ is defined by the stochastic differential equation[UO30; Gil96a]:

$$\frac{dx}{dt} = -\frac{x}{t_{rel}} + \sqrt{D}\,\Gamma(t), \tag{2.9}$$

where t_{rel} is the relaxation time of the process and D is the diffusion coefficient. The time evolution of the process $x(t+\tau)$ from the initial condition $x(t)$ in an interval τ as a function of the particular realization of the white noise $\Gamma(t)$ is given by:

$$x(t+\tau) = x(t)e^{-\frac{\tau}{t_{rel}}} + \sqrt{D}\int_0^\tau dt'\,\Gamma(t+t')\,e^{-\frac{\tau-t'}{t_{rel}}}. \tag{2.10}$$

Since the noise realization is not influenced by previous values of the process, $\langle\Gamma(t+t')|x(t)\rangle = 0$ with $t' > 0$, then the conditional average of the evolution is given by $\langle x(t+\tau)|x(t)\rangle = x(t)\,e^{-\frac{\tau}{t_{rel}}}$, and t_{rel} is effectively the relaxation time of the process.

The conditional variance σ_τ^2 on the evolution in an interval τ is calculated using the white noise property $\langle\Gamma(t+t')\Gamma(t+t'')\rangle = \delta(t'-t'')$ and the formal solution (Eq.2.10) as:

$$\sigma_\tau^2 \equiv \left\langle \left(x(t+\tau) - x(t)\,e^{-\frac{\tau}{t_{rel}}}\right)^2 |x(t)\right\rangle =$$
$$= De^{-\frac{2\tau}{t_{rel}}}\int_0^\tau\int_0^\tau dt'dt''\,\langle\Gamma(t+t')\Gamma(t+t'')\rangle\,e^{\frac{t'+t''}{t_{rel}}} = D\tfrac{t_{rel}}{2}(1-e^{-\frac{2\tau}{t_{rel}}}). \tag{2.11}$$

Since white noise is uncorrelated and symmetrically distributed around zero, expectation values of the type $\left\langle\Gamma(t+t')\Gamma(t+t'')\Gamma(t+t''')\,...\,\Gamma(t+t^{(n)})\right\rangle$ are different from zero only if the n factors are composed of only couples, quartets, or in general groups with an even number of white noise variables multiplied at the same time instant. Considering the combinatorics of the possible permutations of white noise variables in the n-th moment of the conditional evolution distribution we obtain:

$$\left\langle \left(x(t+\tau) - x(t)\,e^{-\frac{\tau}{t_{rel}}}\right)^n |x(t)\right\rangle = \begin{cases} 0, & \text{for odd } n. \\ 1\cdot 3\cdot 5\cdot ...\cdot(n-1)\cdot\sigma_\tau^n, & \text{for even } n. \end{cases} \tag{2.12}$$

This is exactly the form of a Gaussian distribution $N(0,\sigma_\tau^2)$, therefore we can write the probability density of the evolution of the Ornstein-Uhlenbeck process as:

$$p\left(x(t+\tau)|x(t)\right) = N(x(t)e^{-\frac{\tau}{t_{rel}}}, \sigma_\tau^2). \tag{2.13}$$

The steady-state distribution of the process is obtained in the $\tau \to \infty$ limit of no conditioning, that is when the effect of the initial condition is almost surely vanishing because of the time relaxation. The probability density of finding the process in position x with no information on its past history (and also no information on its future states) is then:

$$p(x) = \lim_{\tau\to\infty} p(x(t+\tau) = x|x(t)) = N(0, \sigma_x^2), \tag{2.14}$$

where $\sigma_x^2 = \lim_{\tau\to\infty} \sigma_\tau = D\frac{t_{rel}}{2}$.

Let us now discuss the expectations of past values of the process $x(t-\tau)$ at the time instants $t-\tau$ (with $\tau > 0$) based on the knowledge of the state at time t, $x(t)$. The condition $x(t)$ conveys information on previous states of the process because the dynamics is almost surely continuous and has a finite relaxation time. The conditional expectation of past values is given by:

$$\langle x(t-\tau)|x(t)\rangle = \frac{1}{Z_\tau}\int_{-\infty}^{+\infty} dx\ p(x)\ p(x(t)|x(t-\tau) = x)\ x, \tag{2.15}$$

where the normalization factor is $Z_\tau = \int dx\ p(x)p\left(x(t)|x(t-\tau) = x\right) = \int dx(t-\tau)\ p\left(x(t-\tau), x(t)\right) = p(x(t))$, and it could as well been derived with the Bayes rule. Here and in the following, the integrals are over the whole real line $(-\infty, +\infty)$ when not explicitly written.

In order to estimate Eq.2.15, and also in the analytical calculation of many information-theoretic quantities that we will introduce in Chap.3, we will make use of the following Gaussian integrals (with $A > 0$ and B real numbers):

$$\int dx\ e^{-Ax^2+Bx} = \sqrt{\frac{\pi}{A}}\ e^{\frac{B^2}{4A}}. \tag{2.16}$$

$$\int dx\ xe^{-Ax^2+Bx} = \frac{B}{2A}\sqrt{\frac{\pi}{A}}\ e^{\frac{B^2}{4A}}. \tag{2.17}$$

$$\int dx\ x^2e^{-Ax^2+Bx} = \frac{2A+B^2}{4A^2}\sqrt{\frac{\pi}{A}}\ e^{\frac{B^2}{4A}}. \tag{2.18}$$

Using the result of Gaussian integrals (Eq.2.16) and the probability density for the evolution of the process (Eq.2.13) we obtain the mean and variance of the conditional distribution of past values of the process:

$$\langle x(t-\tau)|x(t)\rangle = x(t)e^{-\frac{\tau}{t_{rel}}} = \langle x(t-\tau)|x(t)\rangle \tag{2.19}$$

$$\sigma_{-\tau}^2 \equiv \left\langle x^2(t-\tau)|x(t)\right\rangle - \langle x(t-\tau)|x(t)\rangle^2 = D\frac{t_{rel}}{2}(1-e^{-\frac{2\tau}{t_{rel}}}) = \sigma_\tau^2. \tag{2.20}$$

We also calculated a few more moments of the distribution considering higher order Gaussian integrals to verify that the conditional distribution of past states is Gaussian and, since the first two moments are equal to the conditional distribution of future states(Eq.2.19), these two distributions are equivalent. The *time symmetry* is a fundamental property of the Ornstein-Uhlenbeck process, and it will imply thermodynamic reversibility as we will discuss in Chapter 5.

2.4 Spectral analysis of stochastic processes

Let us introduce the study of stochastic processes in the frequency domain, that is particularly relevant for the description of oscillating systems. The spectral theory is widely used in electrical engineering[BB86], and it gives tools for analytical calculations of covariance matrices in linear systems.

Let us define the truncated Fourier transform $\hat{x}(w)$ of a trajectory x in the time interval $[0, T]$:

$$\hat{x}_T(w) = \int_0^T dt\ e^{-iwt} x(t), \tag{2.21}$$

where i is the imaginary unit. Let us recall the Euler's formula $e^{i\alpha} = \cos(\alpha) + i\sin(\alpha)$, for any real α.

The Fourier transform $\hat{x}(w)$ is the $T \to \infty$ limit of the truncated Fourier transform, $\hat{x}(w) = \lim_{T\to\infty} \hat{x}_T(w)$. The Fourier transform $\hat{x}(w)$ is a function of the angular frequency w, which is proportional to the ordinary frequency, $f = \frac{w}{2\pi}$. The Fourier transform $\hat{x}(w)$ does not generally converge in the limit $T \to \infty$ for stationary stochastic processes, and its importance is based on the convergence of a related quantity called power spectral density. The power spectral density (PSD) $\mu_{xx}(w)$ of the stochastic process that generates the trajectories $x(t)$ is defined as[Kra+18]:

$$\mu_{xx}(w) = \lim_{T\to\infty} \frac{\langle |\hat{x}_T(w)|^2 \rangle}{T}. \tag{2.22}$$

The PSD $\mu_{xx}(w)$ is a property of the stochastic process and is defined as an ensemble average. It describes the expected spectral content of trajectories, that is the distribution of their power over frequency.

Let us derive an important relation for the Fourier transform of the time derivative of a process[BB86]:

$$\begin{aligned} &\tfrac{1}{\sqrt{T}}(\widehat{\tfrac{dx}{dt}})_T(w) = \tfrac{1}{\sqrt{T}} \textstyle\int_0^T dt\ e^{-iwt} \tfrac{dx}{dt}(t) = \\ &= \tfrac{1}{\sqrt{T}} \left([x(t) e^{-iwt}]_0^T + iw\hat{x}(w) \right) \approx \tfrac{iw\hat{x}(w)}{\sqrt{T}}, \end{aligned} \tag{2.23}$$

where in the second passage we used partial integration, and in the limit $T \to \infty$ the term $[x(t)e^{-iwt}]_0^T$ is almost surely negligible compared to $iw\hat{x}(w)$.

2.4.1 The Wiener-Khinchin-Einstein theorem

Let us define the (non normalized) autocorrelation function $C_{xx}(t,\tau)$:

$$C_{xx}(t,\tau) \equiv C(x(t), x(t+\tau)) \equiv \langle x^*(t)x(t+\tau)\rangle\,, \tag{2.24}$$

where the sign * denotes complex conjugation. The correlation is independent of time for stationary processes, $C_{xx}(t,\tau) = C_{xx}(\tau)$, and can be estimated from a single trajectory in ergodic processes.

Let us write explicitly the squared modulus of the Fourier transform:

$$\begin{aligned}
\langle|\hat{x}_T(w)|^2\rangle &= \left\langle \int_0^T \int_0^T dtdt'\; x^*(t)x(t')e^{-iw(t'-t)}\right\rangle = \\
&= \int_0^T \int_0^T dtdt'\;\; \langle x^*(t)x(t')\rangle\, e^{-iw(t'-t)} = \int_0^T dt \int_{-t}^{T-t} d\tau\; C_{xx}(\tau)e^{-iw\tau} = \\
&= \int_{-T}^0 d\tau\; C_{xx}(\tau)e^{-iw\tau} \int_{-\tau}^T dt + \int_0^T d\tau\; C_{xx}(\tau)e^{-iw\tau} \int_0^{T-\tau} dt = \\
&= \int_{-T}^T d\tau\; C_{xx}(\tau)e^{-iw\tau}(T-|\tau|),
\end{aligned} \tag{2.25}$$

where we made the change of variables $t' \to \tau = t' - t$, and then changed the order of the integrals and splitted the integration region in the two subregions described respectively by $\tau < 0$ and $\tau > 0$. In the third passage we used the stationarity of the autocorrelation, $\partial_t \langle x^*(t)x(t+\tau)\rangle = 0$.

We use this last expression in the PSD definition (Eq.6.7), and we assume that the autocorrelation $C_{xx}(\tau)$ is an integrable function. The term $\frac{|\tau|}{T}C_{xx}(\tau)e^{-iw\tau}$ converges almost everywhere to 0 in the limit $T \to \infty$, and is everywhere bounded by the integrable function $|C_{xx}(\tau)|$, therefore it does not contribute to the integral according to the Lebesgue's dominated convergence theorem[Ber+98]. With this we derived the **Wiener-Khinchin-Einstein theorem**[VK92; Wie30; Khi34; RW99] relating the autocorrelation of a process with its power spectral density:

$$\mu_{xx}(w) = \int_{-\infty}^{+\infty} d\tau\; C_{xx}(\tau)e^{-iw\tau}. \tag{2.26}$$

We see that the correlation function is the Fourier transform of the power spectrum. This relation can be used for the analytical derivation of the correlation function in

linear stationary processes. This is done considering the ensemble average of the squared modulus of the Fourier transform:

$$\begin{aligned}\langle \hat{x}(w)\hat{x}^*(w')\rangle = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} dtdt'\; e^{-iwt+iw't'}\,\langle x(t)x^*(t')\rangle = \\ = \int_{-\infty}^{+\infty} dt'\; e^{i(w'-w)t'}\int_{-\infty}^{+\infty} d\tau\; e^{-iw\tau}C_{xx}(\tau) = \mu_{xx}(w)\int_{-\infty}^{+\infty} dt'\; e^{i(w'-w)t'} = \\ = 2\pi\delta(w'-w)\mu_{xx}(w), \end{aligned} \tag{2.27}$$

where in the last passage we used the integral representation of the Dirac delta distribution.

Let us consider the Ornstein-Uhlenbeck process (2.3) in the white noise representation (Eq.2.9) and let us take the Fourier transform of a particular realizazion:

$$(iw + \frac{1}{t_{rel}})\hat{x}(w) = \sqrt{D}\;\hat{\Gamma}(w), \tag{2.28}$$

where we used the property $\frac{\hat{dx}}{dt}(w) = iw\hat{x}(w)$ calculated in Eq.2.23.

Let us first consider the expectation of the squared modulus of the Fourier transform of white noise:

$$\begin{aligned}\left\langle \hat{\Gamma}(w)\hat{\Gamma}^*(w')\right\rangle = \int\int dtdt'\;\langle\Gamma(t)\Gamma^*(t')\rangle\, e^{-iwt+iw't'} = \\ = \int\int dtdt'\;\delta(t-t')e^{-iwt+iw't'} = \int dt\; e^{-i(w-w')t} = 2\pi\delta(w-w'), \end{aligned} \tag{2.29}$$

where we used the white noise property $\langle\Gamma(t)\Gamma^*(t')\rangle = \delta(t-t')$, and the integral representation of the Dirac delta.

Then we can calculate the left hand side of Eq.2.27:

$$\begin{aligned}\langle \hat{x}(w)\hat{x}^*(w')\rangle = \frac{D}{(\frac{1}{t_{rel}}+iw)(\frac{1}{t_{rel}}-iw')}\left\langle \hat{\Gamma}(w)\hat{\Gamma}^*(w')\right\rangle = \\ = \frac{D2\pi}{\frac{1}{t_{rel}^2}+w^2}\delta(w-w'), \end{aligned} \tag{2.30}$$

Using Eq.2.27 the power spectral density is then:

$$\mu_{xx}(w) = \frac{D}{\frac{1}{t_{rel}^2}+w^2}. \tag{2.31}$$

Taking the inverse transform of the Wiener-Khinchin-Einstein theorem we recover the autocorrelation of the OU process: $C_{xx}(\tau) = \int dw\;\mu_{xx}(w)e^{iw\tau} = \frac{Dt_{rel}}{2}e^{-\frac{\tau}{t_{rel}}}$.

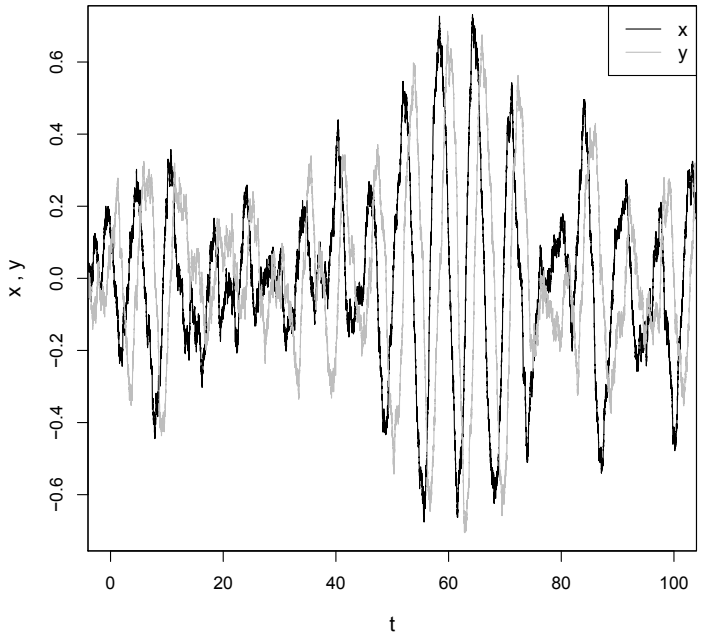


Fig. 2.1: Stochastic dynamics of the negative feedback loop. The parameters are $\beta = 0.1$, $\alpha = 1$, and $D = 0.1$.

2.4.2 Stochastic linear negative feedback loop

We will now apply the spectral analysis to derive the correlation matrix in the linear negative feedback loop model defined by the bidimensional SDE:

$$\begin{cases} \frac{dx}{dt} = -\beta x - \alpha y + \sqrt{D}\ \Gamma_x(t) \\ \frac{dy}{dt} = \alpha x - \beta y + \sqrt{D}\ \Gamma_y(t) \end{cases} \tag{2.32}$$

where β is the damping rate, D is the noise intensity, and α is the interaction parameter. The randomness introduced by the Brownian terms Γ_x and Γ_y is transformed by Eq.2.32 into noisy oscillations whose frequency fluctuates around α.

The autocorrelation of the variables is found with the power spectral density method (Eq.2.27), and like the dynamics is oscillating, $C_{xx}(\tau) = \frac{D}{\beta} e^{-\beta\tau} \cos(\alpha\tau)$. Such result was used in [Wes+09] to test the validity of the stochastic linear negative feedback loop model for the description of circadian oscillations in single cells.

Now we wish to find the cross-correlation $C_{xy}(\tau) \equiv \langle x^*(t)y(t+\tau)\rangle$. Let us first define the cross-spectral density:

$$\mu_{xy}(w) = \lim_{T\to\infty} \frac{\langle \hat{x}_T^*(w)\hat{y}_T(w)\rangle}{T}. \tag{2.33}$$

In analogy to the Wiener-Khinchin-Einstein theorem (Eq.2.26), it is easy to prove the cross-spectral theorem:

$$\mu_{xy}(w) = \int_{-\infty}^{+\infty} d\tau\ C_{xy}(\tau)e^{-iw\tau}. \tag{2.34}$$

The cross-spectral density method, in analogy with Eq.2.27, is written $\langle \hat{y}(w)\hat{x}^*(w')\rangle = 2\pi\delta(w'-w)\mu_{xy}(w)$. Note that in the case of real valued stationary processes it holds $C_{xy}(\tau) = C_{yx}(-\tau)$. Let us then transform the negative feedback loop equations (Eq.2.32) and calculate:

$$\langle \hat{y}(w)\hat{x}^*(w')\rangle = \frac{-4\pi\ i\ D\alpha w\ \delta(w-w')}{(\alpha^2+\beta^2-w^2)^2+4w^2\beta^2}, \tag{2.35}$$

where we used Eq.2.29 and the white noise property $\langle \Gamma_y(t)\Gamma_x^*(t')\rangle = \delta_{xy}\delta(t-t')$. Then we identify the cross-spectral density $\mu_{xy}(w) = \frac{-i\ 2D\alpha w}{(\alpha^2+\beta^2-w^2)^2+4w^2\beta^2}$, and we can calculate the cross-correlation with the inverse transform of Eq.2.34:

$$C_{xy}(\tau) = \int_{-\infty}^{+\infty} dw\ \mu_{xy}(w)e^{iw\tau} = \frac{D}{\beta}e^{-\beta\tau}\sin(\alpha\tau). \tag{2.36}$$

The integral has been calculated with the residue theorem, which is a standard tool in complex analysis[Ahl53; Ber+98] for the integration of real functions whose analytic continuation is vanishing sufficiently fast with w in the complex plane, precisely with $|f(w)|\cdot w \to 0$, and this is the case for $\mu_{xy}(w)$. For such functions the integral on the real axis is equal to the contour integral on the boundary of the upper complex half plane, the $Im(w) > 0$ contribution vanishing when $|w| \to \infty$. The residue theorem then substitutes the contour integral with the sum of residues, which are quantities describing the function asymptotics near to its singularities. The residue on a single pole w_1 (like the singularities of $\mu_{xy}(w)$ in the stochastic linear negative feedback loop model) for a function $f(w)$ is written $Res(w_1) = lim_{w\to w_1}(f(w)*(w-w_1))$. In our case the poles in the upper complex half plane are $w_1 = \alpha+i\beta$ and $w_2 = -\alpha+i\beta$, and their residues are respectively proportional to $e^{(-\beta+i\alpha)\tau}$ and $-e^{(-\beta-i\alpha)\tau}$, then summing we get the functional form $e^{-\beta\tau}\sin(\alpha\tau)$. Interestingly, the correlation analysis describes the linear stochastic negative feedback loop dynamics as a noisy sine-cosine pair. A realization of the dynamics is plotted in Fig.2.1.

2.4.3 Stochastic resonance

Stochastic resonance[Ben+81; Gam+98] is a phenomenon of increased sensitivity to small perturbations due to the insertion of a noise source. This effect is found in nonlinear systems characterized by a form of threshold, like in bistable or excitable systems. Stochastic resonance was experimentally verified in lasers[McN+88], mechanoreceptor neurons[Dou+93], and many more examples. We will not consider stochastic resonance in so much detail here because we will not consider it explicitly in the main results of the present thesis. We will use stochastic resonance in the estimation of the mutual irreversibility in circadian perturbed oscillations, to overcome the spatial sampling discretization. The only model we will consider with a potential stochastic resonance is the receptor-ligand model. Indeed for large Hill coefficients the receptor-ligand system is effectively bistable.

Let us just qualitatively discuss the simplest understood mechanism here, that is the overdamped motion of a Brownian particle in a bistable potential subject to periodic forcing. This can be described by the SDE:

$$\frac{dx}{dt} = -\frac{dV(x)}{dx} + A_0 \cos(\Omega t) + \sqrt{2D}\,\Gamma_t, \tag{2.37}$$

where the bistable potential has the form $V(x) = -\frac{x^2}{2} + \frac{x^4}{4}$, where ΔV is the height of the potential barrier. Transitions between the two local equilibrium states happens with the Kramer's rate $r_K = \frac{1}{\sqrt{2\pi}} e^{-\frac{\Delta V}{D}}$. For small amplitudes the conditional mean oscillates and synchronizes to the periodic forcing with a phase lag:

$$\langle x(t) \rangle = \langle x(t + \frac{2\pi}{\Omega}) \rangle = \bar{x} \cos\Big(\Lambda t - \bar{\phi}\Big) \tag{2.38}$$

The amplitude $\bar{x}$ is a function of the noise intensity, and has a peak at $D_{SR} > 0$ fixed by the trascendental equation $4r_K^2(D_{SR}) = \Omega^2(\frac{\Delta V}{D_{SR}} - 1)$.

Intuitively, when the noise intensity is too small $D << D_{SR}$ then it is too rare to cross the potential barrier, while for too large noise intensities $D >> D_{SR}$ there are many switching for every period of the forcing and therefore no synchronization. The optimal $D = D_{SR}$ correspond to a noise intensity that allows the desired amplification of the signal.

2.5 Ito and Stratonovich interpretations of SDE

The stochastic differential equations that we consider in this thesis have noise terms in the form of Brownian motion dW, or equivalently of white noise Γ. In general

the noise can be multiplicative[Bro+97; San+82], meaning that a function of the state, $g(x(t))$, is the coefficient for the noise intensity, like in $dx = g(x)dW(t)$ (or $\frac{dx}{dt} = g(x)\Gamma(t)$ in the white noise representation). Let us consider more in general g to be an adapted stochastic process, that is a functional of the trajectory up to the time instant at which g is considered, $g(t) = \int_{-\infty}^{t} dt'\, G(x(t'), t', t)$. The calculation of the evolution of the process x in a finite time interval T involves a stochastic integral for which more than one interpretation is possible:

$$I(T) \equiv \int_0^T g(t)dW(t). \tag{2.39}$$

We first describe the Ito interpretation of Eq.2.39. If $g(t)$ is a simple process, meaning that it is constant on time subintervals $[t_j, t_{j+1})$ for a particular partition $(t_0 = 0, t_1, t_2, ..., t_n = T)$, then the Ito integral is defined as:

$$\int_0^T g(t)dW(t) = \sum_{j=0}^{n-1} g(t_j)(W(t_{j+1}) - W(t_j)). \tag{2.40}$$

For a general integrand $g(t)$ being an adapted stochastic process, we need to define a sequence of simple processes $g_n(t)$ such that it converges to the continuous $g(t)$ in the limit $n \to \infty$, $\lim_{n\to\infty} \left\langle \int_0^T |g_n(t) - g(t)|^2 dt \right\rangle = 0$, where the ensemble average is due to the trajectory dependence of $g(t)$. Let us further assume the squared integrability condition of $g(t)$, $\left\langle \int_0^T g^2(t)dt \right\rangle < \infty$. Then, the Ito integral for general integrands is defined as the $n \to \infty$ limit of the Ito integral for simple integrands (Eq.2.40).

The Ito integral is a martingale because of the zero-expectation independent increments of Brownian motion, $\left\langle \int_0^T g(t)dW(t) \right\rangle = 0$, and its quadratic variation is almost surely equal to $[I, I](T) = \int_0^T g^2(t)dt$.

The nonzero quadratic variation of Brownian motion and the Ito scheme of evaluating the integrand value always at the beginning of the subinterval makes the Ito calculus rules to be different from the ones of ordinary calculus. As an example the integral $\int_0^T W(t)dW(t) = \frac{1}{2}W^2(T) - \frac{1}{2}T$ gives the additional term $-\frac{1}{2}T$ in comparison to the ordinary calculus result $\int_0^T g(t)dg(t) = \int_0^T g(t)\frac{dg(t)}{dt}dt = \frac{1}{2}g^2(T)$ with $g(0) = W(0) = 0$. This is seen from the definition Eq.2.40 with $g(t) = W(t)$ and the notation $\Delta_j \equiv W(t_{j+1}) - W(t_j)$:

$$\begin{aligned}\int_0^T W(t)dW(t) &= \lim_{n\to\infty} \sum_{j=0}^{n-1} W(t_j)\Delta_j = \lim_{n\to\infty} \sum_{j=0}^{n-1} \Delta_j \sum_{i=0}^{j-1} \Delta_i = \\ &= \tfrac{1}{2} \lim_{n\to\infty} \left(\sum_{j=0}^{n-1} \sum_{i=0}^{n-1} \Delta_j\Delta_i - \sum_{j=0}^{n-1} \Delta_j^2 \right) = \\ &= \tfrac{1}{2}W^2(T) - \tfrac{1}{2}[W, W](T) = \tfrac{1}{2}W^2(T) - \tfrac{1}{2}T,\end{aligned} \tag{2.41}$$

where in the second passage we used the symmetry of the integration region and in the last passage one should recognise the quadratic variation of Brownian motion that we calculated in Eq.2.5. The effect of Ito calculus is seen in the diagonal homogeneous terms we had to subtract when extending the summation region, these summing up to $-\frac{1}{2}T$ in the limit $n \to \infty$. Thanks to this additional term $-\frac{1}{2}T$ the integral is a martingale $\left\langle \int_0^T W(t)dW(t) \right\rangle = \frac{1}{2}\langle W^2(T)\rangle - \frac{1}{2}T = 0$.

Another effect of the nonzero quadratic variation of Brownian motion and Ito calculus is the evolution of the functions of Brownian motion $f(T, W(T))$. This is described by the **Ito-Doeblin formula** for a function $f(t, x)$ with continuous second partial derivatives as:

$$\begin{aligned} f(T, W(T)) &= f(0, W(0)) + \int_0^T f_t(t, W(t))dt + \\ &+ \int_0^T f_x(t, W(t))dW(t) + \frac{1}{2}\int_0^T f_{xx}(t, W(t))dt, \end{aligned} \tag{2.42}$$

where the dt in the last integral comes from the formal relation $dW^2 = dt$, which is true only under the integral sign and is related to the convergence of the quadratic variation. The Ito-Doeblin formula can be readily generalized for a function of a general stochastic process $f(T, X(T))$ considering the non-vanishing terms up to the second order $dX\ dX$.

A different interpretation of the stochastic integral $I(T)$ is given by Stratonovich that evaluates the integrand in the mid point of each subinterval:

$$\int_0^T g(t)dW(t) = \lim_{|\Pi|\to 0} \sum_{j=0}^{n-1} g(\frac{t_j + t_{j+1}}{2})(W(t_{j+1}) - W(t_j)), \tag{2.43}$$

where $|\Pi| \to 0$ means the limit of vanishing partition intervals as in the Ito integral.

The ordinary rules of calculus apply to the Stratonovich integral, which therefore results not to be a martingale, $\left\langle \int_0^T W(t)dW(t) \right\rangle = \frac{1}{2}\langle W^2(T)\rangle = \frac{1}{2}T$. While physicists generally use the Stratonovich scheme, the Ito calculus is mostly used in quantitative finance where the function $g(t)$ correspond to an asset position. In the algorithmic trading framework the Stratonovich scheme is inappropriate because it would amount to an effective possibility of arbitrage since the choice of investing at time t is influenced by the asset price W at a future time $t + \Delta t$.

Nevertheless there exist an exact correspondence for formulating the same model in the Ito and Stratonovich SDE formulations. The Stratonovich SDE $dx = \alpha(x)dt + \beta(x)dW(t)$ is equivalent to the Ito SDE $dx = (\alpha(x) + \frac{1}{2}\beta(x)\frac{\partial\beta(x)}{dx})dt + \beta(x)dW(t)$. We see that in the case of non-multiplicative noise, $\frac{\partial\beta(x)}{dx} = 0$, the two interpretations

are equivalent. We will use multiplicative noise in our biological models, and we will express them in the Ito formalism. The way we will introduce fluctuations in these models is related to the geometric Brownian motion that we discuss in the following section.

2.5.1 Geometric Brownian motion

The geometric Brownian motion (GBM) is defined with the following SDE in the Ito interpretation:

$$dx = \alpha x dt + \sigma x dW. \tag{2.44}$$

α is called *drift* and $\sigma > 0$ is called *volatility*. The multiplicative noise xdW induces fluctuations of a magnitude comparable with the process state at each time. Considering that in the Ito interpretation $d\ln x = \frac{dx}{x} - \frac{dxdx}{2x^2} = \frac{dx}{x} - \frac{\sigma^2 dt}{2}$ according to the Ito-Doeblin formula (2.42), the evolution of GBM is written as a function of the Brownian motion evolution $W(t)$:

$$x(t) = x(0)\ e^{\sigma W(t) + (\alpha - \frac{\sigma^2}{2})t}. \tag{2.45}$$

If $\alpha = 0$, the GBM is a martingale, and this gives an important feature to the Log-Normal distribution describing the stochastic evolution: the dynamics almost surely vanishes (such property is valid for any $\alpha < \frac{\sigma^2}{2}$). In the Stratonovich representation the GBM (Eq.2.44) is written $dx = (\alpha - \frac{\sigma^2}{2})\ x\ dt + \sigma\ x\ dW$.

2.6 Fokker-Planck equation

Stochastic differential equations can in general be expressed in terms of partial differential equations (PDEs), while the opposite is not true. This transformation is precisely described by the Feynman-Kac theorem[Shr12]. Here we describe a special case of this, that is the transformation of a SDE into a PDE involving the evolution probability $P(x(t+\tau)|x(t))$. Let us begin with the SDE:

$$dx = \beta(x(t), t)\ dt + \sigma(x(t), t)\ dW, \tag{2.46}$$

where the coefficients of drift $\beta(x(t), t)$ and diffusion $\sigma(x(t), t)$ can be explicitly dependent on both the process $x(t)$ and the time instant t.

Let $h(x)$ be a deterministic function of the stochastic process $x(t)$, but not explicitly of time, and assume it to have continuous derivatives. Then its variation is given

by the Ito-Doeblin formula (Eq.2.42) in the differential form (without the partial derivative with respect to time):

$$dh(x(t)) = \partial_x h(x(t))\ dx + \tfrac{1}{2}\partial_{xx} h(x(t))\ dx\ dx = \\ = \partial_x h(x(t))\ (\beta(x(t),t)\ dt + \sigma(x(t),t)\ dW) + \tfrac{1}{2}\partial_{xx} h(x(t))\ \sigma^2(x(t),t)\ dt. \quad (2.47)$$

Now we integrate this equation from t to $t+\tau$ and take conditional expectations given the initial condition for the dynamics $x(t)$ at time t:

$$\int_{-\infty}^{+\infty} dx(t+\tau)\ h(x(t+\tau)) P(x(t+\tau)|x(t)) = \\ = h(x(t)) + \int_0^\tau dt' \int_{-\infty}^{+\infty} dx(t+t')\ \partial_x h(x(t+t'))\ \beta(x(t+t'),t+t')\ P(x(t+t')|x(t)) + \\ + \int_0^\tau dt' \int_{-\infty}^{+\infty} dx(t+t')\ \tfrac{1}{2}\partial_{xx} h(x(t+t'))\ \sigma^2(x(t+t'),t+t')\ P(x(t+t')|x(t)). \quad (2.48)$$

We integrate by parts with respect to $dx(t+t')$ with the assumptions $\partial_x h(x(t+t'))\ \beta(x(t+t'),t)\ P(x(t+t')|x(t)) \to 0$ and $\frac{1}{2}\partial_{xx} h(x(t+t'))\ \sigma^2(x(t+t'),t+t')\ P(x(t+t')|x(t)) \to 0$ in the boundary limits $x(t+t') \to \pm\infty$, and then we take the partial derivative with respect to τ obtaining:

$$\int_{-\infty}^{+\infty} h(x(t+\tau)) \Big(\tfrac{\partial}{\partial\tau} P(x(t+\tau)|x(t)) + \tfrac{\partial}{\partial x(t+\tau)} [\beta(x(t+\tau),\tau) P(x(t+\tau)|x(t))] - \\ - \tfrac{1}{2} \tfrac{\partial^2}{\partial x^2(t+\tau)} [\sigma^2(x(t+\tau),\tau) P(x(t+\tau)|x(t))] \Big)\ dx(t+\tau) = 0. \quad (2.49)$$

Since the last expression should hold for any function $h(x)$, then using the fundamental lemma for the calculus of variations we obtain the **Fokker-Planck equation** corresponding to the SDE in Eq.2.46:

$$\tfrac{\partial}{\partial\tau} P(x(t+\tau)|x(t)) = -\tfrac{\partial}{\partial x(t+\tau)} [\beta(x(t+\tau),\tau) P(x(t+\tau)|x(t))] + \\ + \tfrac{1}{2} \tfrac{\partial^2}{\partial x^2(t+\tau)} [\sigma^2(x(t+\tau),\tau) P(x(t+\tau)|x(t))], \quad (2.50)$$

where the two terms on the RHS correspond respectively to drift and diffusion. This equation is often referred to as forward Kolmogorov equation.

Similarly we can find a PDE with partial derivatives with respect to the condition $x(t)$, for a generic evolution time interval $\tau > 0$, and it is called the backward Kolmogorov equation:

$$-\tfrac{\partial}{\partial t} P(x(t+\tau)|x(t)) = \beta(x(t),t) \tfrac{\partial}{\partial x(t)} P(x(t+\tau)|x(t)) + \\ + \tfrac{1}{2}\sigma^2(x(t),t) \tfrac{\partial^2}{\partial x^2(t)} P(x(t+\tau)|x(t)). \quad (2.51)$$

2.7 Path integrals

The probability of any particular trajectory of a stochastic process is the probability of its corresponding noise source realization[CD01]. Let us consider a one-dimensional stochastic process (with constant diffusion coefficient) whose dynamics is influenced by an externally controlled parameter λ:

$$\dot{x} = f(x,\lambda) + \sqrt{D}\,\Gamma_t. \tag{2.52}$$

This is equivalent in the Ito representation to $dx = f(x,\lambda)dt + \sqrt{D}\,dW_t$, and in the Stratonovich representation to $dx = f(x+\frac{dx}{2},\lambda)dt + \sqrt{D}\,dW_t$. Let us start writing the probability of a particular trajectory x_0^t in the interval $[0,t]$ in the Ito representation. This is a function of the corresponding realization of the stochastic component of the dynamics that is the noise Γ_t. This is done discretizing the time interval in n steps, therefore considering Brownian increments $W(\frac{t(k+1)}{n}) - W(\frac{tk}{n})$ in the place of Γ_t, and then taking the limit $n \to \infty$:

$$\begin{aligned}
p(x_0^t|x(0)) &= \lim_{n\to\infty} D^{-\frac{n}{2}} \prod_{k=0}^{n-1} \frac{e^{-\frac{\left(W(\frac{t(k+1)}{n}) - W(\frac{tk}{n})\right)^2}{2\frac{t}{n}}}}{\sqrt{2\pi\frac{t}{n}}} = \\
&= \lim_{n\to\infty} \prod_{k=0}^{n-1} \frac{e^{-\frac{\left(x(\frac{t(k+1)}{n}) - x(\frac{tk}{n}) - f(x(\frac{tk}{n}),\lambda)\frac{t}{n}\right)^2}{2D\frac{t}{n}}}}{\sqrt{2\pi D\frac{t}{n}}} = \\
&= \lim_{n\to\infty} (\tfrac{n}{2\pi Dt})^{\frac{n}{2}} e^{-\frac{n}{2Dt}\sum_{k=0}^{n-1}\left(x(\frac{t(k+1)}{n}) - x(\frac{tk}{n}) - f(x(\frac{tk}{n}),\lambda)\frac{t}{n}\right)^2} = \\
&= \lim_{n\to\infty} (\tfrac{n}{2\pi Dt})^{\frac{n}{2}} e^{-\frac{1}{2D}\sum_{k=0}^{n-1}\left(\frac{x(\frac{t(k+1)}{n}) - x(\frac{tk}{n})}{\frac{t}{n}} - f(x(\frac{tk}{n}),\lambda)\right)^2\frac{t}{n}} = \\
&= \lim_{n\to\infty} (\tfrac{n}{2\pi Dt})^{\frac{n}{2}} e^{-\frac{1}{2D}\int_0^t dt'\ (\dot{x} - f(x,\lambda))^2} = \frac{1}{Z} e^{-\frac{1}{2D}\int_0^t dt'\ (\dot{x} - f(x,\lambda))^2},
\end{aligned} \tag{2.53}$$

where the probability density is expressed in the space of trajectories that is described by the differential $dx_0^t \equiv \prod_{k=0}^{n-1} d(x(\frac{t(k+1)}{n}) - x(\frac{tk}{n}))$. The term $D^{-\frac{n}{2}}$ in the first passage is the Jacobian of the transformation from Brownian space dW_0^t to trajectory space dx_0^t, $|\prod_{k=0}^{n-1} \frac{\partial(x(\frac{t(k+1)}{n}) - x(\frac{tk}{n}))}{\partial(W(\frac{t(k+1)}{n}) - W(\frac{tk}{n}))}|^{-1} = D^{-\frac{n}{2}}$, where the off-diagonal terms vanish assuming that $f(x,\lambda)$ is smooth. Note that, importantly, the divergent normalization factor $Z \equiv \lim_{n\to\infty} (\frac{n}{2\pi Dt})^{\frac{n}{2}}$ is independent of the trajectory. We will therefore be able to compare the probability density of different trajectories just considering the relation $p(x_0^t|x(0)) \propto e^{-\frac{1}{2D}\int_0^t dt'\ (\dot{x} - f(x,\lambda))^2}$.

With the Stratonovich discretization scheme the Jacobian is $\left| \prod_{k=0}^{n-1} \frac{\partial(x(\frac{t(k+1)}{n}) - x(\frac{tk}{n}))}{\partial(W(\frac{t(k+1)}{n}) - W(\frac{tk}{n}))} \right|^{-1} = D^{-\frac{n}{2}} \prod_{k=0}^{n-1} (1 - \frac{1}{2} \frac{\partial f(x,\lambda)}{\partial x}|_{\frac{tk}{n}} \frac{t}{n}) \to \lim_{n\to\infty} D^{-\frac{n}{2}} e^{-\frac{1}{2}\int_0^t dt' \, \frac{\partial f(x,\lambda)}{\partial x}}$. The Stratonovich path integral is therefore:

$$p(x_0^t|x(0)) \propto e^{-\int_0^t dt' \, \left(\frac{1}{2D}(\dot{x} - f(x,\lambda))^2 + \frac{1}{2}\frac{\partial f(x,\lambda)}{\partial x}\right)}. \tag{2.54}$$

The trajectory probability density has the form $p(x_0^t|x(0)) \propto e^{-S(x,\dot{x},\lambda)}$, where $S(x, \dot{x}, \lambda)$ is the so called *Onsager-Machlup action functional*[OM53; MO53]. An analogous expression is found for general multidimensional Langevin processes with multiplicative noise[Che+06]. We will consider bidimensional processes (x, y) with path densities $\hat{p}(x_0^t|y_0^t, x(0))$, where y_0^t is considered as a fixed trajectory in the action functional even if x has an influence on y in the dynamics. Therefore the path density $\hat{p}(x_0^t|y_0^t, x(0))$ is not a conditional probability, and we will use the symbol $\hat{p}$ to denote such path densities. The physicist use of path integrals in the Stratonovich interpretation rather than Ito was recently justified within the supersymmetric theory of stochastics[Ovc16; MW14].

3 Information thermodynamics on bipartite systems

3.1 Shannon entropy

Shannon entropy is the most commonly used measure of uncertainty and unpredictability [CT12]. Its definition is inspired by the process of taking a choice and communicating the information about this choice[Sha01]. Let us consider a situation in which we ask a question in a precise way so that only a number n of answers are possible. One of these answers is chosen, the process being described by *a priori* probabilities $p_1, p_2, \ ... \ , p_n$. The Shannon entropy $H(p_1, p_2, \ ... \ , p_n)$ measures the amount of information that we get once we are said which of the n possible answers to the question is chosen. Its functional form is:

$$H(p_1, p_2, \ ... \ , p_n) \equiv -K \sum_{i=1}^{n} p_i \log p_i, \tag{3.1}$$

where K is a positive constant.

$H(p_1, \ ... \ , p_n)$ is the only functional form satisfying the following assumptions:

-$H(p_1, \ ... \ , p_n)$ is continuous in the p_is.
-if all the answers are equally likely, that is $p_i = \frac{1}{n} \ \ \forall i$, then $H(p_1, \ ... \ , p_n)$ is a monotonic increasing function of n.
-if the choice can be decomposed into two successive choices, the first choice between the two subsets $[p_1, ..., p_m]$ and $[p_{m+1}, ..., p_n]$ (with $n > m \geq 1$) and the second choice within the subsets, then $H(p_1, \ ... \ , p_n)$ has the branching property: $H(p_1, \ ... \ , p_n) = H(p_1 + p_2 + .. + p_m, p_{m+1} + p_{m+2} + ... + p_n) + (p_1 + p_2 + .. + p_m)H(p_1, ..., p_m) + (p_{m+1} + p_{m+2} + ... + p_n)H(p_{m+1}, ..., p_n)$.

The constant K and the base of the logarithm specify the unit measure for entropy. We will use $K = 1$ and the natural logarithm, then the entropy is measured in natural units of information $[Nats]$. In the case of only two possible events described by probabilities p_1 and p_2 the maximum amount of entropy is obtained when the two probabilities are equal $p_1 = p_2 = \frac{1}{2}$, while entropy is vanishing in the limits $p_1 \to 0$ ($p_2 \to 1$) and $p_1 \to 1$ ($p_2 \to 0$).

The thermodynamics entropy S is recovered in the microcanonical ensemble from Eq.3.1 choosing $K = K_B$, where K_B is the Boltzmann constant. The microcanonical ensemble is a statistical description of thermodynamics that gives a microscopic interpretation and derivation of the macroscopic classical experiments with gas at low pressure[Hua87]. The entropy is there defined as being proportional to the number n of microstates compatible with the macroscopic observable internal energy, $S = K_B \ln n$. These microstates are defined to be all equiprobable, $p_i = \frac{1}{n}$, then Eq.3.1 reduces to $H = K_B \ln n$, which is the Boltzmann thermodynamic entropy $H = S$. Like the Boltzmann constant, the unit measure for entropy is Joules per Kelvin, $\frac{[J]}{[K]} = \frac{[Kg \cdot m^2]}{[K \cdot s^2]}$.

While in Eq.3.1 the space of events is discrete, we have to consider the entropy of continuous variables because we will be dealing with stochastic differential equations. The differential entropy of a random variable X described by the distribution $p(X = x) = p(x)$ is defined as:

$$H(X) = -\int_{-\infty}^{+\infty} dx\ p(x) \ln p(x). \tag{3.2}$$

Note that, contrary to its discrete states counterpart, the differential entropy (Eq.3.3) admits negative values. This is because the differential entropy considers densities rather than actual probabilities and is not the continuous limit of Eq.3.1. In fact the term $-\ln dx$ diverges in the $dx \to 0$ limit of the discrete states entropy.

In the following discussion of dynamic linear stochastic models we will repeatedly consider the entropy of Gaussian distributions $N(\mu, \sigma^2)$. $H(N(\mu, \sigma^2))$ is calculated with the Gaussian integrals (Eq.2.16) as:

$$H(N(\mu, \sigma^2)) = \ln\left(\sigma\sqrt{2\pi e}\right). \tag{3.3}$$

Entropy can be defined for multidimensional variables as well. Consider the two continuous random variables X and Y and their joint distribution $p(x, y)$. Their joint entropy is defined as:

$$H(X, Y) = -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} dxdy\ p(x, y) \ln p(x, y). \tag{3.4}$$

The conditional entropy $H(Y|X)$ measures the average uncertainty in the variable Y when the value of variable X is known. It is defined as:

$$H(Y|X) = H(X, Y) - H(X) = -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} dxdy\ p(x, y) \ln p(y|x). \tag{3.5}$$

3.1.1 Mutual information and transfer entropy

The mutual information[CT12] between two variables X and Y, named $I(X,Y)$, is defined as the average reduction in uncertainty on the value of variable Y that occurs once we get to know the value of variable X. Its natural definition in terms of entropies is therefore $I(X,Y) = H(Y) - H(Y|X)$:

$$I(X,Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} dxdy\; p(x,y) \ln \frac{p(x,y)}{p(x)p(y)}. \tag{3.6}$$

From Eq.3.6 we see that the mutual information is a symmetric measure, $I(X,Y) = I(Y,X)$. Let us write the mutual information as a thermal average, meaning in the form of an expectation value over the ensemble probability density $p(x,y)$:

$$I(X,Y) = \left\langle \ln \frac{p(x,y)}{p(x)p(y)} \right\rangle = \left\langle \ln \frac{p(y|x)}{p(y)} \right\rangle = \left\langle I^{st}(X,Y) \right\rangle, \tag{3.7}$$

where we defined the stochastic mutual information $I^{st}(X,Y) \equiv \ln \frac{p(y|x)}{p(y)}$ which depends on the particular realization (x,y).

Let us consider the stochastic dynamics of two interacting variables x and y and let us assume that it is described by a ergodic stationary process as discussed in Chapter 2. Then their joint probability density for a generic time t is defined, $p(x_t, y_t)$. The joint probability density of the states of the variables at two distinct time points, t and $t+\tau$, separated by a time shift τ, is also defined, $p(x_t, y_t, x_{t+\tau}, y_{t+\tau})$, as well as three-elements joint probabilities like $p(x_t, y_t, x_{t+\tau})$.

The conditional mutual information $I(x_t, y_{t+\tau}|y_t)$ is defined as the mutual information between the variables x_t and $y_{t+\tau}$ that exist when the value of variable y_t is known. In this particular form it is generally called, after Schreiber[Sch00], the transfer entropy $T_{x\to y}(\tau)$, because it is the additional amount of information on the evolution of the dynamics of variable y that is given by the knowledge of variable x:

$$\begin{aligned} T_{x\to y}(\tau) = I(x_t, y_{t+\tau}|y_t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} dx_t dy_t dy_{t+\tau}\; p(x_t, y_t, y_{t+\tau}) \ln \frac{p(x_t, y_{t+\tau}|y_t)}{p(x_t|y_t)p(y_{t+\tau}|y_t)} = \\ = \left\langle \ln \frac{p(x_t, y_{t+\tau}|y_t)}{p(x_t|y_t)p(y_{t+\tau}|y_t)} \right\rangle = \left\langle \ln \frac{p(y_{t+\tau}|x_t, y_t)}{p(y_{t+\tau}|y_t)} \right\rangle = \left\langle T^{st}_{x\to y}(\tau) \right\rangle, \end{aligned} \tag{3.8}$$

where $T^{st}_{x\to y}(\tau) \equiv \ln \frac{p(y_{t+\tau}|x_t, y_t)}{p(y_{t+\tau}|y_t)}$ is the stochastic (realization-dependent) counterpart of the transfer entropy.

The transfer entropy $T_{x\to y}(\tau)$ is a measure of information, i.e. reduction of uncertainty, and it is not a proper terminology to call it information flow [Jam+16]). This is because conditioning on x_t in $T^{st}_{x\to y}(\tau)$ does not select the information flow (or

transfer) from variable x to variable y, but only quantifies an increase in predictive power that, most importantly, involves a synergy with previous knowledge, i.e. with y_t. We will discuss more in detail in Chapter 4, while introducing the information decomposition[WB10; Bar15], why the transfer entropy is not a measure of causal influence[Auc+17; Jam+16]. We will still keep the term "information flow" when describing transfer entropies, as it is commonly done in the literature[IS13; Par+15].

Let us also introduce here the backward transfer entropy $T_{x\to y}(-\tau)$, measuring information flow towards past[Ito16; Auc+18] and defined as:

$$\begin{aligned} T_{x\to y}(-\tau) = I(x_t, y_{t-\tau}|y_t) = I(x_{t+\tau}, y_t|y_{t+\tau}) = \\ = \left\langle \ln \frac{p(y_t|x_{t+\tau}, y_{t+\tau})}{p(y_t|y_{t+\tau})} \right\rangle = \left\langle T^{st}_{x\to y}(-\tau) \right\rangle, \end{aligned} \tag{3.9}$$

where we considered time stationarity of the process.

Applications of the transfer entropy range from cellular automata[Liz+08], to neurobiological assessments of consciousness[Lee+15], and it is a recurring quantity in descriptions of information thermodynamics and fluctuation theorems[Par+15; Auc+19b; Auc+18], which is our main interest.

Kullback-Leibler divergence

We introduce here a measure of distance between probability distributions. It has an entropy-like form but it is asymmetric, and is not properly an information measure. We will show in the following sections that its use in stochastic thermodynamics is fundamentally related to dissipation and the II Law[Kaw+07]. Given two one-dimensional probability distributions $p(x)$ and $q(x)$, their Kullback-Leibler divergence $D(p||q)$ is defined as:

$$D(p||q) = \int dx\; p(x) \ln \frac{p(x)}{q(x)}. \tag{3.10}$$

It is always positive, and it vanishes only if $p(x) = g(x)\ \ \forall x$. This is found minimizing $D(p||q)$ varying $q(x)$ with the Lagrangian multipliers method conditioned on the normalization $\int dx\; q(x) = 1$. The generalization of Eq.3.10 to multivariate distributions is straightforward.

The Kullback-Leibler divergence $D(p||q)$ measures the distinguishability between the two probability distributions $p(x)$ and $q(x)$. The Chernoff-Stein lemma[CT12] formalizes this intuition characterizing the probability of incorrect guessing in the sense of hypothesis testing with an observation limited to n random samples. In particular, guessing that they are generated from $p(x)$ when the true distribution

is $q(x)$ happens with a probability that is asymptotically equal to $e^{-nD(p||q)}$ for $n \to \infty$.

3.2 Optimizing information transmission: the small-noise approximation

We now discuss the optimization of the mutual information in the bivariate case $I(x, y)$ (Eq.3.6) with a relevant biological example in mind: the regulation of gene expression by transcription factors. This was indeed the motivating problem that led to the introduction of the small-noise approximation by Tkačik, Callan and Bialek [Tka+08b].

Let us call x the concentration of a transcription factor (TF) which regulates the expression of a single gene whose concentration is called y. Let us assume that the variation of the signal x is slow enough for the response y not to be influenced by previous values of x, this being the steady-state assumption. When the value of the signal is fixed to a particular x, the response y is still a Random variable because of the stochasticity of chemical reactions[Rei+18], like the mRNA production and degradation, and of the interaction between TFs and the promoter region of the gene on the DNA. The physics of the regulatory element is summarized by the conditional distribution $P(y|x)$. The noise in transcriptional regulation is described by the variance of $P(y|x)$, here called $\sigma^2_{y|x}(x)$ and explicitly dependent on the signal x, meaning that different concentrations of TF lead to different values of uncertainty in gene expression. $\sigma^2_{y|x}(x)$ can be so high that looking at the TF concentration one might only be able to roughly distinguish between two transcriptional states, gene "ON" and gene "OFF", this corresponding to $1\ bit$ of information. In general, the logarithm of the number of states of the output y that one is effectively able to discriminate varying the value of the input x is described by the mutual information $I(x, y)$ (Eq.3.6). Given the channel property $P(y|x)$, the output distribution $P(y) = \int dx P(x, y)$ and the mutual information $I(x, y)$ are determined by the distribution of inputs $P(x)$. Now we wish to optimize the information transmission from the concentration of transcription factors to the gene expression level for a given channel $P(y|x)$, that means finding the input distribution $P(x)$ that leads to the highest possible value of the mutual information, $I^*(x, y)$. Such maximal value $I^*(x, y)$ is called channel capacity[CT12]. Searching for the optimal input distribution $P^*(x)$ is a really difficult task in general, but efficient numerical algorithms are available[Tka+08a; Bla72].

If the noise in transcriptional regulation is small, then the small-noise approximation (SNA)[Tka+08b] becomes useful. The SNA amounts to considering the mutual

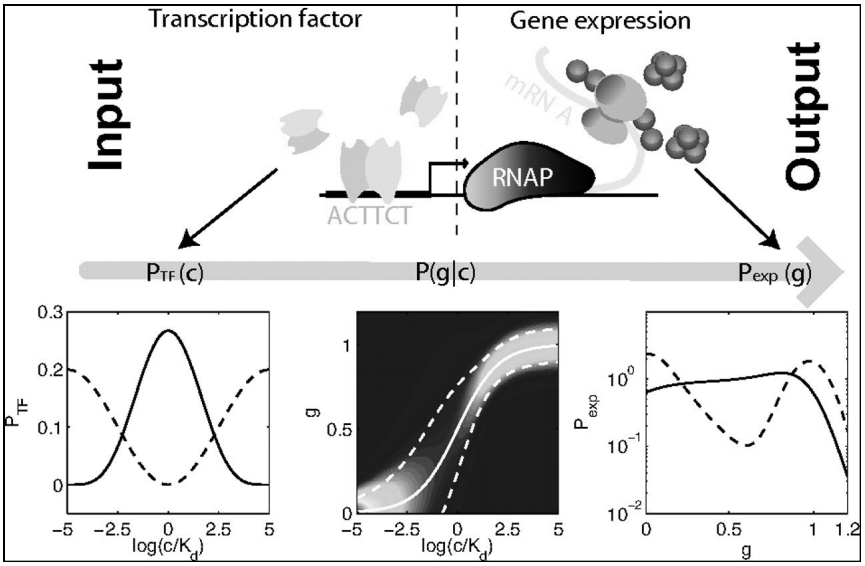


Fig. 3.1: Input-output relations in transcriptional regulation. The concentration of transcription factors $P_{TF}(c)$ is passed through the noisy channel $P(g|c)$ to give the output distribution of gene expression $P_{exp}(g)$. In the lower panels two different input distributions result in two different output distributions. The plot is taken from Ref.[Tka+08b]. Copyright (2008) National Academy of Sciences, U.S.A.

information in the form $I(x,y) = \int dx\ P(x) \int dy\ P(y|x) \ln P(y|x) - \int dy P(y) \ln P(y)$ and approximating the distribution of outputs $P(y)$ with the distribution that the cell would generate in the absence of transcriptional noise $P_{SNA}(y)$:

$$\begin{aligned} P_{SNA}(y) &= \int dx\ P(x)\ \delta(y - \langle y|x\rangle) = \\ &= P(x = x_{SNA}(y))\ |\tfrac{d\langle y|x\rangle}{dx}|^{-1}_{x=x_{SNA}(y)}, \end{aligned} \tag{3.11}$$

where $x_{SNA}(y)$ is the value of the input whose conditional output average $\langle y|x\rangle$ is equal to y, and we assumed the monotonicity of $\langle y|x\rangle$ (see Fig.3.1).

Let us further assume that the steady-state noise of the regulatory element is Gaussian:

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma^2_{y|x}(x)}} e^{-\frac{(y-\langle y|x\rangle)^2}{2\sigma^2_{y|x}(x)}}, \tag{3.12}$$

where $\sigma^2_{y|x}(x) = \langle (y - \langle y|x\rangle)^2 \rangle\,|_x$ is the conditional variance at a particular transcription factor concentration x. With the SNA (Eq.3.11) the mutual information reduces to:

$$\begin{aligned} I_{SNA}(x,y) &= -\int dy\ P_{SNA}(y) \ln P_{SNA}(y) \\ &-\tfrac{1}{2} \int dy\ P_{SNA}(y)\ \ln\left(2\pi e\ \sigma^2_{y|x}(x_{SNA}(y))\right). \end{aligned} \tag{3.13}$$

Note that $dx\ P(x) = dy\ P_{SNA}(y)$ is a change of variables and not an approximation. In the calculation of the mutual information $I(x,y) = H(y) - H(y|x)$, the SNA is just approximating the output entropy $H(y)$ with the noiseless channel output entropy $H_{SNA}(y)$, and in the conditional entropy $H(y|x)$ is approximating the channel to be Gaussian. The SNA makes it possible to analytically solve the optimization problem $\max_{P_{SNA}(y)} I(x,y)$ subject to the probability constraint $\int dy\ P_{SNA}(y) = 1$. Using the method of Lagrangian multipliers[Ber14], we have to maximize the augmented functional $J = I(x,y) - \lambda(\int dy\ P_{SNA}(y)\ - 1)$, where λ is the Lagrangian multiplier. Taking the functional derivative of J with respect to $P_{SNA}(y)$ and setting it to 0 we obtain:

$$\begin{aligned} 0 = \tfrac{\delta J}{\delta P_{SNA}(y)}(P^*_{SNA}(y)) = \\ = -\ln P^*_{SNA}(y) - 1 - \tfrac{1}{2}\ln\left(2\pi e\ \sigma^2_{y|x}(x_{SNA}(y))\right) - \lambda. \end{aligned} \tag{3.14}$$

where $P^*_{SNA}(y)$ is the optimal distribution of average responses. From the last expression we derived the form $P^*_{SNA}(y) = \frac{1}{Z}\ \frac{1}{\sigma^2_{y|x}(x_{SNA}(y))}$, where Z is a function of the Lagrangian multiplier λ and has to be determined with the normalization constraint, $Z = \int \frac{dy}{\sigma^2_{y|x}(x_{SNA}(y))}$. Using Eq.3.11 we can write this result in terms of the optimal input distribution $P^*_{SNA}(x)$, and this was our original aim:

$$P^*_{SNA}(x) = \frac{1}{Z}\ \frac{1}{\sigma^2_{y|x}(x)}\ |\frac{d\,\langle y|x\rangle}{dx}|_x. \tag{3.15}$$

The SNA optimal input distribution $P^*_{SNA}(x)$ (Eq.3.15) is driven by two factors: the preferential use of reliable input values, meaning those that have a smaller conditional output variance $\sigma^2_{y|x}(x)$, and the preferential use of the dynamic regime, meaning those input values that correspond to the steepest part of the conditional expectation function $\langle y|x\rangle$, those with larger $|\frac{d\langle y|x\rangle}{dx}|$. Substituting $P^*_{SNA}(x)$ into Eq.3.13 we get the optimal mutual information $I_{SNA}(x,y) = \ln\frac{Z}{\sqrt{2\pi e}}$. We see that $Z = \int \frac{dy}{\sigma^2_{y|x}(x_{SNA}(y))}$ can be estimated from data on the conditional variance, where the condition on the signal is equivalently expressed as a condition on the mean expression level $x_{SNA}(y)$ thanks to the SNA (Eq.3.11). Therefore in the small noise regime we can estimate the channel capacity just looking at gene expression fluctuations at steady-state, regardless of the input-output relation structure $\langle y|x\rangle$.

Application to the pattern formation in Drosophila embryo

Cells in the developing fruit fly *Drosophila* embryo need to have information on where they are located within the embryo in order to differentiate accordingly. Such positional information is stored in the *Hunchback* (Hb) gene response y to the maternally established concentration pattern of the *Bicoid* (Bcd) transcription

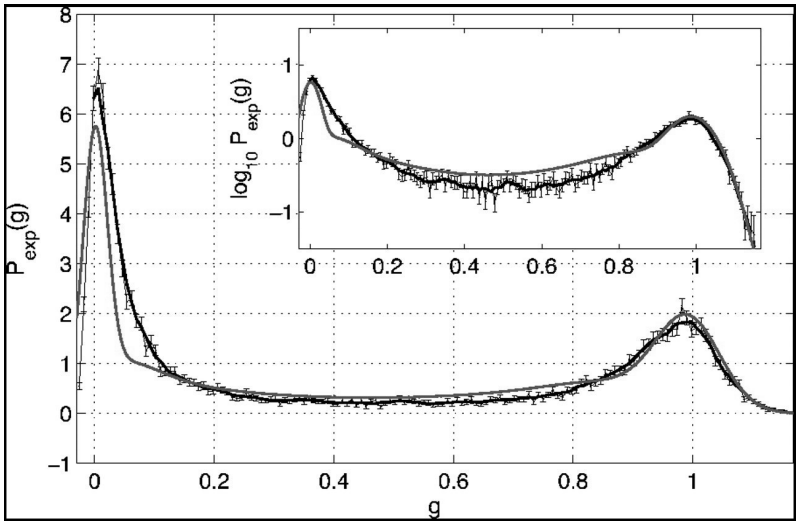


Fig. 3.2: The gray continuous line is the parameter-free prediction of the *Hunchback* gene expression distribution based on the principle of maximum information transmission from *Bicoid* transcription factors, and it is compared with experimental measures (black bars and lines) on Drosophila embryos. The plot is taken from Ref.[Tka+08b]. Copyright (2008) National Academy of Sciences, U.S.A.

factor x. Here the probability of inputs $P(x)$ is the fraction of (nearby) cells with Bcd concentration equal to x, while the conditional $P(y|x)$ describes fluctuations of Hb expression levels among cells which are subject to the same Bcd concentration, these cells being also spatially close for an efficient transmission of the positional information in a noisy environment. Hb and Bcd concentrations have been measured simultaneously *in vivo* in each cell of several *Drosophila* embryos [Gre+07], thus obtaining an estimate of the joint probability $P(x, y)$. Using the estimated $P(x, y)$ in Eq.3.6 for each embryo replica they obtain the information transmitted and it is equal to $I(x, y) = 1.5 \pm 0.15\ bits$. This value is close to the theoretical maximum of $I^*(x, y) = 1.8\ bits$, which is calculated numerically from the experimental noise variance $\sigma^2_{y|x}(x_{SNA}(y))$ with only the Gaussian assumption and no SNA ($x_{SNA}(y)$ is here just a deterministic relation between average gene expression and transcription factor concentration). This means that the generated gene expression distribution $P(y)$ should be such that the dynamic regime of the input-output relation $\langle y|x\rangle$ is used to obtain intermediate levels of Hb expression and transmit more than one bit of information. In Fig.3.2 the experimentally measured distribution of Hb gene expression $P(y) = P_{exp}(g)$ is shown to be in good agreement with the optimal $P^*(y)$.

Summarizing, the authors of [Tka+08b]-[Dub+13] were able to predict a nontrivial feature of transcriptional regulation in *Drosophila* embryo development, that is the use of intermediate levels of gene expression to reliably encode three transcriptional states with a pattern of TF concentration along the embryo, in contrast to the simple

1 bit gene "ON/OFF" model. Most importantly, the prediction was based on the hypothesis that information transmission is maximized in the process to cope with the limitations imposed by the randomness of chemical reactions and interactions.

3.3 Jarzynski-Crooks nonequilibrium thermodynamics

3.3.1 Work and heat dissipation in nonequilibrium Markovian systems

Let us consider a physical system whose state at time t is described by the variable $\vec{x_t}$, and whose dynamics is influenced by a parameter λ_t which can be controlled by an external agent. As an example, in the classical gas-piston model the $\vec{x_t}$ represents the positions and momenta of all particles, while λ_t corresponds to the volume to which the gas is confined, such volume being controlled by the piston. Let the system evolve under Langevin dynamics, so that probability densities are associated to particular realizations.

Let us consider a fixed protocol for the control parameter λ_t, that is a fixed sequence of states $\lambda_0, \lambda_1, \ldots, \lambda_N$ corresponding to the time intervals $[t_0, t_1), [t_1, t_2), \ldots, [t_{N-1}, t_N]$. The dynamics of the system is stochastic, and evolves under the influence of the λ protocol. A particular realization is written:

$$\vec{x_0} \xrightarrow{\lambda_1} \vec{x_1} \xrightarrow{\lambda_2} \vec{x_2} \xrightarrow{\lambda_3} \ldots \vec{x_{N-1}} \xrightarrow{\lambda_N} \vec{x_N}. \tag{3.16}$$

Here we begin the discussion of fluctuation theorems with the Crooks derivation[Cro98] of an exact expression for the dissipated work in microscopically reversible systems. Such expression is called the detailed fluctuation theorem because it holds for single realizations of the process. We write the derivation in discrete time and space for the simplicity of notation, but it can be readily generalized to continuous time and space.

At equilibrium (for a single fixed value of $\lambda_t = \lambda \ \forall t$) the ensemble of trajectories is canonically distributed. The probability of observing a state $\vec{A}$ with energy $E(\vec{A}, \lambda)$ is given by:

$$P(\vec{A}|\lambda) = \frac{e^{-\beta E(\vec{A},\lambda)}}{\sum_i e^{-\beta E(i,\lambda)}} = e^{\beta\left(F(\beta,\lambda) - E(\vec{A},\lambda)\right)}, \tag{3.17}$$

where $\beta = K_B T$ and $F = -\frac{1}{\beta}\sum_i e^{-\beta E(i,\lambda)}$ is the free energy of the system.

The control protocol is such that λ_t moves with discontinuous jumps. These instantaneous variations $\lambda_t \to \lambda_{t+1}$ correspond to a change in energy $E(\vec{x_t}, \lambda_t) \to E(\vec{x_t}, \lambda_{t+1})$ that is interpreted as mechanical work performed on the system:

$$W_t = E(\vec{x_t}, \lambda_{t+1}) - E(\vec{x_t}, \lambda_t). \tag{3.18}$$

Then, after the work parameter has jumped, the system evolves in the finite time interval $[t, t+1)$ absorbing an amount of heat given by $Q_t = E(\vec{x_{t+1}}, \lambda_{t+1}) - E(\vec{x_t}, \lambda_{t+1})$. The **I Law of thermodynamics** holds for the internal energy variation:

$$W_t + Q_t = E(\vec{x_{t+1}}, \lambda_{t+1}) - E(\vec{x_t}, \lambda_t) \equiv \Delta E_t. \tag{3.19}$$

The reversible work W_r in switching between two configurations λ_0 and λ_1 of the external parameter is defined as the free energy difference between the two corresponding equilibrium ensembles, $W_r \equiv \Delta F = F(\beta, \lambda_1) - F(\beta, \lambda_0)$. This corresponds to the amount of work performed on the system when the variation of the work parameter is enough slow for the system to be always in the canonical equilibrium state corresponding to the instantaneous λ. Such a process is then said to be reversible.

The dissipative work W_d is defined as the difference between the actual work performed on the system and the corresponding reversible work, $W_d = W - W_r$. Importantly, it depends on the particular realization of the dynamics of the system $(\vec{x_0}, \vec{x_1}, ..., \vec{x_N})$ and on that of the control parameter $\lambda_0, \lambda_1, \ ... \ , \lambda_N$. Note that the work performed over the whole process W is just the sum of the discrete steps, $W = \sum_{t=0}^{N-1} W_t$.

Let us now consider the *backward path* corresponding to Eq.3.16:

$$\vec{x_N} \xrightarrow{\lambda_N} \vec{x_{N-1}} \xrightarrow{\lambda_{N-1}} \vec{x_{N-2}} \xrightarrow{\lambda_{N-2}} \ ... \ \vec{x_1} \xrightarrow{\lambda_1} \vec{x_0} \tag{3.20}$$

Let us assume the system to be *Markovian*, so that the evolution depends only on the present state of the system and not on its history. Then the probability of a path under a fixed protocol can be written as:

$$\begin{aligned} P\left(\vec{x_0} \xrightarrow{\lambda_1} \vec{x_1} \xrightarrow{\lambda_2} \vec{x_2} \xrightarrow{\lambda_3} \ ... \ \vec{x_{N-1}} \xrightarrow{\lambda_N} \vec{x_N}\right) = \\ = P\left(\vec{x_0} \xrightarrow{\lambda_1} \vec{x_1}\right) \cdot P\left(\vec{x_1} \xrightarrow{\lambda_2} \vec{x_2}\right) \cdot \ ... \ \cdot P\left(\vec{x_{N-1}} \xrightarrow{\lambda_N} \vec{x_N}\right) \end{aligned} \tag{3.21}$$

The dynamics in phase-space is supposed to be *microscopically reversible*:

$$\frac{P\left(\vec{A} \xrightarrow{\lambda} \vec{B}\right)}{P\left(\vec{B} \xrightarrow{\lambda} \vec{A}\right)} = \frac{P(B|\lambda)}{P(A|\lambda)} = e^{-\beta(E(B,\lambda)-E(A,\lambda))}. \tag{3.22}$$

Given the properties of microscopic reversibility (Eq.3.22), Markovian dynamics (Eq.3.21), and the expression for the work performed in single jumps of the control parameter (Eq.3.18), it is obtained an expression for the **dissipative work** on a single realization of the dynamics:

$$W_d = \frac{1}{\beta} \ln \left(\frac{P(\vec{x_0}|\lambda_0) \cdot P\left(\vec{x_0} \xrightarrow{\lambda_1} \vec{x_1} \xrightarrow{\lambda_2} \vec{x_2} \xrightarrow{\lambda_3} \ ... \ \vec{x_{N-1}} \xrightarrow{\lambda_N} \vec{x_N}\right)}{P(\vec{x_N}|\lambda_N) \cdot P\left(\vec{x_N} \xrightarrow{\lambda_N} \vec{x_{N-1}} \xrightarrow{\lambda_{N-1}} \vec{x_{N-2}} \xrightarrow{\lambda_{N-2}} \ ... \ \vec{x_1} \xrightarrow{\lambda_1} \vec{x_0}\right)} \right) \tag{3.23}$$

The dissipated work W_d results to be a function of the probability ratio of observing the particular path $(\vec{x_1}, \ ... \ , \vec{x_N})$ starting from an equilibrium configuration with λ_0 and of observing the time-reversed conjugate path $(\vec{x_N}, \ ... \ , \vec{x_1})$ starting from an equilibrium configuration with λ_N and with the time-reversed protocol. This fundamental relation between the time irreversibility of paths and the dissipative work is called the detailed fluctuation theorem.

3.3.2 The Jarzynski nonequilibrium equality for free energy differences

Using Eq.3.23 and the normalization of backward paths it follows the **Jarzynski's Nonequilibrium work theorem**[Jar97]:

$$\left\langle e^{-\beta W} \right\rangle = \sum_{\vec{x_0}, \ ... \ ,\vec{x_N}} P(\vec{x_0}|\lambda_0) \cdot P\left(\vec{x_0} \xrightarrow{\lambda_1} \vec{x_1} \xrightarrow{\lambda_2} \vec{x_2} \xrightarrow{\lambda_3} \ ... \ \vec{x_{N-1}} \xrightarrow{\lambda_N} \vec{x_N}\right) e^{-\beta W} =$$

$$= e^{-\beta \Delta F}. \tag{3.24}$$

In the original formulation of Jarzynski[Jar97] this exponential average was derived for (deterministic) Hamiltonian trajectories in phase space starting from a canonical distribution, with the Hamiltonian being a function of the control protocol $\lambda_F(t)$ and therefore time-dependent.

In general the variation of the control parameter from the initial state λ_0 to the final state λ_N is performed over a finite time interval $t_N - t_0 < \infty$, and it can be made of jumps, or it can be everywhere continuous (for $N \to \infty$). In both cases the ensemble of trajectories cannot be totally relaxed to the steady-state distribution described by the value of λ at each time point, and the system is said to evolve *out*

of equilibrium. The more the λ protocol is fast, the more the ensemble of trajectories will be far from the instantaneous steady-state distributions during the process. As a consequence the work W performed on these nonequilibrium ensembles is on average larger than the free energy difference between the two steady-state configurations $\Delta F = F(\beta, \lambda_1) - F(\beta, \lambda_0)$. This is obtained from the Jarzynski equality (Eq.3.24), considering the convexity of the exponential. We recall the Jensen's inequality for a convex function f, namely $\langle f(x) \rangle \geq f(\langle x \rangle)$. Then the **II Law of Thermodynamics** is written:

$$\langle W \rangle \geq \Delta F. \tag{3.25}$$

The Jarzynski equality (Eq.3.24) is therefore a generalization of the II Law of Thermodynamics (Eq.3.25). Indeed in single realizations we can observe apparent violations $W < \Delta F$, while the II Law (Eq.3.25) is valid at the macroscopic level, i.e. on the ensemble average. Fluctuations of work values around the free energy difference of equilibrium states ΔF have been shown to be relevant in small thermodynamic systems like molecular motors[Sei12]. The Jarzynski equality relates an equilibrium quantity, that is the free energy difference between equilibrium configurations (which are described by canonical distributions), to a nonequilibrium quantity, that is the exponential average of work values over many realizations of the dynamics under a control protocol that can bring the system arbitrarily out of equilibrium depending on how fast the external parameter is varied.

3.3.3 Applicability and rare realizations

The Jarzynski nonequilibrium work theorem (Eq.3.24) suggests an experimental method to estimate the free energy difference between two equilibrium configurations by means of work measurements in many realizations of the same nonequilibrium experiment, such experiment being described by the protocol $(\lambda_0, \lambda_1, ..., \lambda_N)$. The estimated free energy difference ΔF_{est} between the two equilibrium states corresponding to λ_N and λ_0 over n_{rep} replicas of the same nonequilibrium experiment is given by:

$$\Delta F_{est} = -\frac{1}{\beta} \ln \left(\frac{1}{n_{rep}} \sum_{j=1}^{n_{rep}} e^{-\beta W_j} \right), \tag{3.26}$$

where the W_js are the different values of work performed on the system obtained in the different replicas.

We would like to be able to use the approximation $\Delta F \approx \Delta F_{est}$. In practice this average is dominated by very rare realizations[LG05] and the convergence is really slow with the number of replicas n_{rep}. We discuss here a characterization that Jarzynski gave of these rare dominant realizations and an estimate of the num-

ber of realizations needed for convergence[Jar06a], that is relevant for numerical simulations and experimental works.

Let us now consider the continuous-time limit of the dynamics, that in a finite time path of duration T is the limit of increasingly many subintervals of vanishing time lenght. Such continuous trajectories are in general described by stochastic differential equations (see Chapter 2). We denote a trajectory defined in $[0, T]$ with $x_{[0,T]}(t)$. Then the dissipative work in the system trajectory $x_{[0,T]}(t)$ under the control protocol (also in continuous time) $\lambda_F(t)$ is given by:

$$W_d(x_{[0,T]}(t)) = \beta^{-1} \ln \left(\frac{p_{\lambda_F}(x_{[0,T]}(t))}{p_{\lambda_R}(x_{[0,T]}(t))} \right), \tag{3.27}$$

where p_{λ_F} denotes a probability density in the space of trajectories under the control protocol λ_F, $x_{[0,T]}(t)$ is the time-reverse conjugate of the original trajectory $x_{[0,T]}(t)$, $x_{[0,T]}(t) = x_{[0,T]}(T - t)$, and λ_R is the time reversal of the forward control protocol, $\lambda_R(t) = \lambda_F(T-t)$. Equation (3.27), that was derived by Crooks[Cro98] for stochastic Markovian dynamics with microscopic reversibility as described in the last section, it was shown by Jarzynski[Jar06a] to be also valid for deterministic Hamiltonian dynamics as already mentioned.

Let us now consider the expected value of the dissipated work from Eq.3.27:

$$\beta \left\langle W_d(x_{[0,T]}(t)) \right\rangle = \left\langle \ln \left(\frac{p_{\lambda_F}(x_{[0,T]}(t))}{p_{\lambda_R}(x_{[0,T]}(t))} \right) \right\rangle, \tag{3.28}$$

where the brakets indicate the average over the ensemble of realizations under the forward protocol λ_F. This formula is known as the KPB relation[Kaw+07; Par+09] after Kawai, Parrondo and Van den Broeck.

An equivalent formulation of the integral fluctuation theorem (Eq.3.24) as a function of the dissipative work $W_d(x_{[0,T]}(t)) = W(x_{[0,T]}(t)) - \Delta F$ is:

$$\left\langle e^{-\beta W_d(x_{[0,T]}(t))} \right\rangle = 1, \tag{3.29}$$

where the dependence of the dissipative work on the particular realization $x_{[0,T]}(t)$ is explicitly written.

If we sample dissipated work values with many replicas of the nonequilibrium experiment under the influence of the forward control protocol $\lambda_F(t)$, most of the

times we obtain values around $W_d(x^{typ}_{[0,T]}(t))$ where $x^{typ}_{[0,T]}(t)$ is the "typical realization" of the process[Jar06a] defined as:

$$x^{typ}_{[0,T]}(t) = \left\langle x_{[0,T]}(t) \right\rangle = \int dx_{[0,T]}(t)\; p_{\lambda_F}(x_{[0,T]}(t))\; x_{[0,T]}(t), \tag{3.30}$$

where the integral is performed over the trajectory space. $x^{typ}_{[0,T]}(t)$ can be seen as an average realization of the dynamics.

The integral in Eq.3.29 is dominated by those rare realizations which are close to the "dominant realization" of the process, which is defined as:

$$x^{dom}_{[0,T]}(t) = \frac{\left\langle x_{[0,T]}(t) e^{-\beta W_d(x_{[0,T]}(t))} \right\rangle}{\left\langle e^{-\beta W_d(x_{[0,T]}(t))} \right\rangle} = \int dx_{[0,T]}(t)\; p_{\lambda_F}\left(x_{[0,T]}(t)\right)\; e^{-\beta W_d(x_{[0,T]}(t))} x_{[0,T]}(t) =$$
$$= \int dx_{[0,T]}(t)\; p_{\lambda_R}\left(x_{[0,T]}(t)\right)\; x_{[0,T]}(t), \tag{3.31}$$

where in the last passage we used the detailed fluctuation theorem Eq.3.27.

From equations 3.30-3.31 we see that, importantly, the dominant realizations of the dynamics with forward protocol λ_F, those that contribute more to the exponential average of Eq.3.29, are the time-reverse conjugate trajectories of the typical realizations of the dynamics under the backward protocol λ_R.

Let us consider the gas-piston model as a practical example of applicability of the Jarzynski nonequilibrium work theorem to the estimation of free energy differences. Lua and Grosberg[LG05] calculated that, in a non-interacting particles model, when the piston is pushed into the gas and in the fast piston limit, the dominant realizations are in the tail of the Maxwell distribution of velocities and correspond to the unlikely event of no particle-piston collisions and therefore no work. These rare events are equivalent to the time-reversed conjugates of the typical trajectories observed in the backward experiment, which is the fast volume expansion. This is illustrated in Fig.3.3.

Similarly, in the backward experiment corresponding to very fast increase of the volume by moving the piston outward, the dominant realizations for the convergence of the exponential average $\left\langle e^{-\beta W_d(x_{[0,T]}(t))} \right\rangle_{\lambda_R}$ (under the backward protocol λ_R) are those in the tail of the Maxwell distribution where all the molecules are moving fast enough to hit the piston while it is moving and perform work. Here with "fast piston" we mean with much greater velocity compared to the thermal particle speed[Hua87] given by $v_{th} = \sqrt{\frac{3}{m\beta}}$. The authors also showed in the gas-piston setting that the average number of replicas of the experiment in order for the exponential average to converge increases exponentially with the system size, making the Jarzynski method

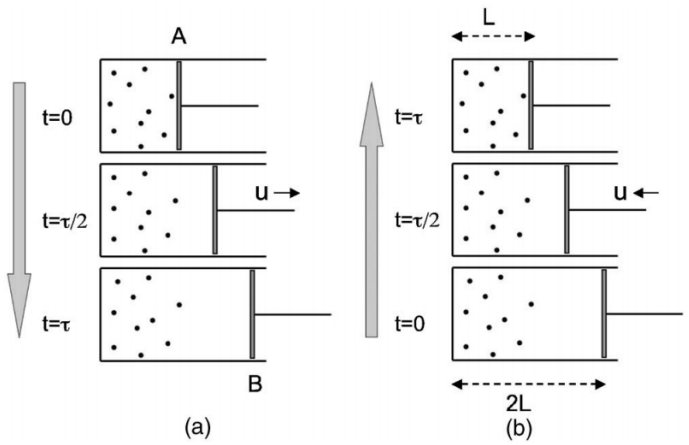


Fig. 3.3: *a)* In a typical backward experiment, that is the gas expansion, if the piston is pulled quickly then the particles have fewer collisions with the piston and perform less work compared to the slower reversible process. The time-reverse conjugates of these typical realizations are the dominant realizations in the forward experiment, the gas compression. *b)* The rare event of no collisions in the volume compression is the dominant realization for the exponential average in the Jarzynski equality, and corresponds to an initial highly asymmetric spatial distribution of the molecules (or of their velocities). The plot is taken from Ref.[Jar06a], DOI: 10.1103/PhysRevE.73.046105. Copyright 2006 by the American Physical Society.

in general unpractical. Nevertheless, if we assume the possibility to characterize the tails of the distribution looking at the shape of the finite-samples histogram around the typical realizations of the dynamics, that is where we normally get samples, then it is reasonable to use the Jarzynski method in the free energy difference estimation.

3.4 Measurement information and feedback control

Measuring observables on a small thermodynamic system is a physical operation that involves an interaction between an observer and the system. In other words some work has to be performed (and also quantified and interpreted) in order to gain information on the system state. This is necessarily the case because the information gained from measurements can be used to extract work from a system in an isothermal cycle, which would be in contradiction to the II Law of thermodynamics if no work is associated with the measurement[SU13]. Let us recall the Kelvin–Planck statement[MS87] of the II Law:

"*It is impossible to devise a cyclically operating device, the sole effect of which is to absorb energy in the form of heat from a single thermal reservoir and to deliver an equivalent amount of work.*"

The idea of challenging the thermodynamics II Law with measurement-feedback control scheme is commonly known as the Maxwell's demon paradox, and was formulated by Szilard with a theoretical single-particle gas-piston information engine[Szi64]. The Szilard thought experiment consist of a single-particle gas-piston model and an intelligent being called Maxwell's demon who is able to measure in which part of the volume the particle is located (the left half or the right half) obtaining 1 bit of information. Then he can rapidly move the piston (without collisions with the particle) to the half volume position thus increasing the free energy of the single-particle gas, and slowly perform an isothermal expansion extracting $K_B T \ln 2$ of work from a single heat bath going back to the initial macroscopic state. The validity of the II Law imply that at least an amount of work or heat dissipation equal to $K_B T \ln 2$ is involved in the measurement process. An information engine inspired by the Maxwell's demon has been experimentally realized with a colloidal particle in an electrical field[Toy+10].

Here we review the general theory of nonequilibrium feedback control as it was formulated by Sagawa and Ueda[SU12; SU10; SU13]. Relating to the previous sections, we basically derive the stochastic (nonequilibrium) thermodynamics of a controlled system whose control protocol is influenced by measurements of the system state at previous time steps.

3.4.1 Measurements and information

We keep the same formalism and discretization scheme as in section 3.3. Let us call $X_n \equiv (\vec{x_0}, \ldots, \vec{x_n})$ the system's discretized trajectory (time series) up to time step n, and $\Lambda_n \equiv (\lambda_0, \ldots, \lambda_n)$ the discretized control protocol which is taken to be constant in the time-continuous dynamics within each of the discretization steps. The dynamics starts with a change in the control parameter $\lambda_0 \to \lambda_1$ and then with the heat absorption at constant λ_1 corresponding to the system's state transition $\vec{x_0} \to \vec{x_1}$. The corresponding backward experiment starts with a state transition $\vec{x_0} \to \vec{x_1}$ instead and then with a change in the control parameter $(\lambda_0^B \to \lambda_1^B) \equiv (\lambda_n \to \lambda_{n-1})$. Such asymmetry is taken in order to have the same length for X and Λ vectors, but it has no effect in the continuous time limit. The backward trajectory and backward control protocol corresponding to X and Λ are denoted respectively as $\widetilde{X}$ and $\widetilde{\Lambda}$, with $\widetilde{\vec{x_n}} = \vec{x_{N-n}}$ and $\widetilde{\lambda_n} = \lambda_{N-n}$. To be precise we should take $\widetilde{\lambda_n}$ to be the time-reversal of λ_{N-n}, denoted λ^*_{N-n}, and for a quantity which is odd under time reversal (like a magnetic field) we would have to change sign $\widetilde{\lambda_n} = \lambda^*_{N-n} = -\lambda_{N-n}$. We

will not consider explicitly this case and assume that λ is even under time reversal not to overload the formalism, but the results we derive are easily generalized to such a case. The same can be said about the system state backward trajectory if its description explicitly involves the momentum p.

Let us now introduce noisy measurements of the system state. At each time step $t_n = n\Delta t$ we perform a measurement whose outcome y_n has an error described by probability $P(y_n|X_n)$. The measurement is Markovian if it can be performed much faster compared to the shorter time scale of the system, $P(y_n|X_n) = P(y_n|\vec{x_n})$. $Y_n = (y_0, \ ... \ , y_n)$ is the vector of measurements up to time step n. We further assume that the measurement itself has no istantaneous perturbation on the system, and its influence results only from the corresponding control protocol $\Lambda_N(Y_{N-1}) \equiv (\lambda_0, \lambda_1(Y_0), \ ... \ , \lambda_N(Y_{N-1}))$. This assumption of measuring without an interaction with the system is in contradiction to the thermodynamics II Law, and it will result in potentially negative values of the entropy production. We also make the reasonable assumption of no explicit correlation between measurements $P(y_k|X_k, Y_{k-1}) = P(y_k|X_k)$, meaning that the error in the measurement process is described by uncorrelated Random noise.

Let us now define the following quantity:

$$P_C(Y_n|X_n) \equiv \prod_{k=0}^{n} P(y_k|X_k). \tag{3.32}$$

In the presence of feedback, that is when the control protocol λ_k at each time step k depends on the outcome of previous measurements Y_{k-1}, the quantity $P_C(Y_n|X_n)$ is different from the conditional distribution $P(Y_n|X_n)$, $P_C(Y_n|X_n) \neq P(Y_n|X_n)$. This is because the knowledge of the system state gives information on previous values of the protocol since that affected its dynamics, then in general $P(y_k|X_n, Y_{k-1}) = P(y_k|X_n) \neq P(y_k|X_k)$ for $k < n$.

The joint distribution of measurements and dynamics with feedback control is given by:

$$\begin{aligned} P(X_n, Y_n) = P(\vec{x_0}, y_0) \textstyle\prod_{k=0}^{n-1} P(y_{k+1}|X_{k+1}) P(\vec{x_{k+1}}|X_k, \lambda_{k+1}(Y_k)) = \\ = P(X_n|\Lambda_N(Y_{n-1})) P_C(Y_n|X_n), \end{aligned} \tag{3.33}$$

where $P(X_n|\Lambda_N(Y_{n-1})) = P(\vec{x_0}) \prod_{k=0}^{n-1} P(\vec{x_{k+1}}|X_k, \lambda_{k+1}(Y_k))$ is the probability density of path X_n under the effect of the **fixed** control protocol $\Lambda_N(Y_{n-1})$. The condition on the whole history up to time k for the evolution of the dynamics, that is $P(\vec{x_{k+1}}|X_k, \lambda_{k+1}(Y_k))$ instead of just $P(\vec{x_{k+1}}|\vec{x_k}, \lambda_{k+1}(Y_k))$, is to consider the case of non-Markovian processes.

We now consider the probability of the corresponding time-reversal trajectory under the **fixed** backward control protocol, meaning the time-reversal control protocol which depends on measurements on the forward process, $\Lambda_N(Y_{n-1})$. Since no feedback is involved in the backward experiment, this probability is simply written for a given initial state $\widetilde{\vec{x_0}} = \vec{x_n}$ as:

$$P(\widetilde{X_n}|\widetilde{\vec{x_0}}, \Lambda_N(Y_{n-1})) = \prod_{k=0}^{n-1} P(\vec{x_{k+1}}|\widetilde{X_k}, \widetilde{\lambda_k}). \tag{3.34}$$

The initial distributions for the forward and backward trajectories initial states $P_{0,F}$ and $P_{0,B}$ don't have to be necessarily canonical, and multiple heat baths can be involved in the process. The entropy production $\varphi(X_N|\Lambda_N(Y_{N-1}))$ along trajectory X_N under the **fixed** control protocol $\Lambda_N(Y_{N-1})$ (for fixed Y_{N-1}) is defined as:

$$\varphi(X_N|\Lambda_N(Y_{N-1})) = -\ln\Big(P_{0,B}(\widetilde{\vec{x_0}}|Y_{N-1})\Big) + \ln(P_{0,F}(\vec{x_0})) - \sum_i \beta_i Q_i(X_N|\Lambda_N(Y_{N-1})), \tag{3.35}$$

where $Q_i(X_N|\Lambda_N(Y_{N-1}))$ is the heat absorbed from the i-th heat bath with inverse temperature β_i.

For a fixed control protocol $\Lambda_N(Y_{N-1})$ without feedback (with fixed Y_{N-1}) the detailed fluctuation theorem holds as we already showed in section 3.3:

$$\varphi(X_N|\Lambda_N(Y_{N-1})) = \frac{P(X_N|\Lambda_N(Y_{N-1}))}{P(\widetilde{X_N}|\Lambda_N(Y_{N-1}))}. \tag{3.36}$$

It is important to note that in Eq.3.36 the conditional probabilities $P(X_N|\Lambda_N(Y_{N-1}))$ and $P(\widetilde{X_N}|\Lambda_N(Y_{N-1}))$ describe experiments where no feedback is performed. This ensures that the detailed fluctuation theorem of Eq.3.36 holds, because the framework is not qualitatively different from the previous section 3.3, and Eq.3.36 is just a generalization of Eq.3.27 to multiple heat baths and non Markovian dynamics. The probability of trajectory X_N under the condition of measuring Y_{N-1} is $P(X_N|Y_{N-1})$, and in general $P(X_N|\Lambda_N(Y_{N-1})) \neq P(X_N|Y_{N-1})$. This inequality is true even without feedback because the information on the dynamics from measurements is different compared to the information on the dynamics from the control protocol influence. Importantly, note that the relation between Y_{N-1} and $\Lambda_N(Y_{N-1})$ is assumed to be always unknown.

This interpretation of Eq.3.36 is different from what is written in [SU12], but we will get to the same fluctuation theorem. There they calculate $P(X_N|\Lambda_N(Y_{N-1}))$ and $P(\widetilde{X_N}|\Lambda_N(Y_{N-1}))$ from the two qualitatively different experiments with and without feedback, and then postulate the detailed fluctuation theorem to hold. In particular they define $P(X_N|\Lambda_N(Y_{N-1}))$ as the probability of observing the time series X_N in the subset of time series that produced measurements Y_{N-1}, that is the actual

conditional probability $P(X_N|Y_{N-1})$. We note that $P(X_N|Y_{N-1})$ is characterized by both measurement information and feedback control influence, therefore the validity of Eq.3.36 with this interpretation is not ensured and the physical interpretation of the entropy production $\varphi(X_N|\Lambda_N(Y_{N-1}))$ would be different. Therefore we do not agree with this interpretation of feedback thermodynamics, and we wonder if authors in [SU12] really intended this interpretation.

The feedback does not play a role in the detailed fluctuation theorem Eq.3.36, but it does in the thermal averages where we have to estimate the joint probability of process and measurements (Eq.3.33). Let us also note that the first value of the control parameter λ_0 (the last in the backward control parameter, $\widetilde{\lambda_N} = \lambda_0$) cannot be influenced by the dynamics.

3.4.2 Fluctuation theorems with feedback control

Following again Sagawa[SU12] and keeping a formalism consistent with the previous section 3.1 we define the stochastic transfer entropy from paths to measurements as $T^{st}_{X_k \to Y_{k-1}} \equiv \ln\left(\frac{P(Y_k|X_k,Y_{k-1})}{P(Y_k|Y_{k-1})}\right)$, and it will play a role in feedback thermodynamics. Here the term Y_{k-1} in $T^{st}_{X_k \to Y_{k-1}}$ comes from the shift from the more standard transfer entropy $T^{st}_{X_k \to Y_k} \equiv \ln\left(\frac{P(Y_{k+1}|X_k,Y_k)}{P(Y_{k+1}|Y_k)}\right)$. The transfer entropy $T_{X_k \to Y_{k-1}} = \left\langle T^{st}_{X_k \to Y_{k-1}} \right\rangle$, defined as the thermal average of its stochastic counterpart, quantifies the information that one gets with the last measurement y_k (done at time step k) on the time series up to step k, X_k, considering the knowlegde that one already had at the previous time step $k-1$ given by previous measurements. We further define the path sum I^{st}_c of the stochastic transfer entropy, and we call it *stochastic measurement information* [SU12]:

$$\begin{aligned} I^{st}_c(Y_n, X_n) \equiv \textstyle\sum_{k=0}^{n} T^{st}_{X_k \to Y_{k-1}} = \sum_{k=0}^{n} \ln\left(\frac{P(Y_k|X_k,Y_{k-1})}{P(Y_k|Y_{k-1})}\right) = \\ = \textstyle\sum_{k=0}^{n} \ln\left(\frac{P(y_k|X_k)}{P(Y_k|Y_{k-1})}\right) = \ln\left(\frac{P_C(Y_n|X_n)}{P(Y_n)}\right). \end{aligned} \tag{3.37}$$

Note that $T^{st}_{X_k \to Y_{k-1}} = \ln\left(\frac{P(Y_k|X_k,Y_{k-1})}{P(Y_k|Y_{k-1})}\right)$ is a forward transfer entropy. Let us now consider the exponential average of $\varphi(X_N|\Lambda_N(Y_{N-1})) + I_c(Y_N, X_N)$, meaning the thermal average over the real dynamics with feedback described by the joint probability $P(X_N, Y_N)$:

$$\begin{aligned} \left\langle e^{-\varphi(X_N|\Lambda_N(Y_{N-1}))-I^{st}_c(Y_N,X_N)} \right\rangle = \left\langle \frac{P(Y_N)P(\widetilde{X_N}|\Lambda_N(Y_{N-1}))}{P_C(Y_N|X_N)P(X_N|\Lambda_N(Y_{N-1}))} \right\rangle = \\ = \int\int dX_N dY_N \; P(Y_N)P(\widetilde{X_N}|\Lambda_N(Y_{N-1})) = 1, \end{aligned} \tag{3.38}$$

where we used Eq.3.37, Eq.3.36, Eq.3.33, and the normalization of probabilities. The integral is performed over the N-steps time series and measurements space.

Eq.3.38 is a generalized form of the Jarzynski integral fluctuation theorem (Eq.3.24), and it describes the stochastic thermodynamics of measurement-feedback control models. Using Jensen's inequality we obtain a *generalized form of the II Law of Thermodynamics* for the entropy production $\Phi \equiv \langle \varphi(X_N, \Lambda_N(Y_{N-1})) \rangle$ involving the measurement information $I_c \equiv \langle I_c(Y_N, X_N) \rangle$:

$$\Phi \geq -I_c. \tag{3.39}$$

Eq.3.39 sets a lower bound to the entropy production that is less than zero, and the case $\Phi < 0$ is made possible by feedback control. The limit equality $\Phi = -I_c$ is obtained if the dynamics has reversibility with feedback control, $P(X_N|Y_N) = P(\widetilde{X_N}|\Lambda_N(Y_{N-1}))$, and therefore the quantity $\varphi(X_N, \Lambda_N(Y_{N-1})) + I_c(Y_N, X_N)$ does not fluctuate. The standard II Law of Thermodynamics, that is $\Phi \geq 0$, is apparently not satisfied just because the entropy production in the feedback controller (the Maxwell's demon) has not been considered. If the Maxwell's demon is considered to be part of the system, then the II Law is satisfied and the paradox is solved[Mar+09].

Let us now define the *feedback efficacy parameter* γ_f as:

$$\gamma_f \equiv \int P(\widetilde{Y_N}|\Lambda_N(Y_{N-1}))\, dY_N, \tag{3.40}$$

where $P(\widetilde{Y_N}|\Lambda_N(Y_{N-1}))$ is the probability density of measuring $\widetilde{Y_N}$ in an experiment with fixed control protocol $\Lambda_N(Y_{N-1})$. Note that, in the presence of feedback, a variation of Y_N in the integral is simultaneously changing the control protocol $\Lambda_N(Y_{N-1})$, then such quantity is not normalized in general, $\gamma_f \neq 1$.

Let us now consider the measurements to be Markovian, namely $P(y_n|X_n) = P(y_n|x_n)$, and therefore also $P_C(\widetilde{Y_N}|\widetilde{X_N}) = P_C(Y_N|X_N)$. In [SU12] they also explicitly require the time-reversed symmetry of measures, $P(y_n^*|x_n^*) = P(y_n|x_n)$, and that is needed if the variable x is a momentum. In this framework the feedback efficacy parameter γ_f is related to the entropy production $\varphi(X_N|\Lambda_N(Y_{N-1}))$ (Eq.3.35-Eq.3.36) by the following integral fluctuation theorem:

$$\begin{aligned}
\left\langle e^{-\varphi(X_N|\Lambda_N(Y_{N-1}))} \right\rangle &= \int\int dX_N dY_N\ P(X_N, Y_N) \tfrac{P(\widetilde{X_N}|\Lambda_N(Y_{N-1}))}{P(X_N|\Lambda_N(Y_{N-1}))} = \\
&= \int\int dX_N dY_N\ P_C(Y_N|X_N) P(\widetilde{X_N}|\Lambda_N(Y_{N-1})) = \\
&= \int\int dX_N dY_N\ P_C(\widetilde{Y_N}|\widetilde{X_N}) P(\widetilde{X_N}|\Lambda_N(Y_{N-1})) = \\
&= \int dY_N\ P(\widetilde{Y_N}|\Lambda_N(Y_{N-1})) = \gamma_f.
\end{aligned} \tag{3.41}$$

Contrary to the entropy production $\Phi \equiv \langle \varphi(X_N|\Lambda_N(Y_{N-1})) \rangle$ that has to be estimated with a thermal average with feedback control, the efficacy parameter

$\gamma_f = \int dY_N \; P(\widetilde{Y_N}|\Lambda_N(Y_{N-1}))$ can be estimated with experiments with fixed control protocol, but in order to do this the full information on how feedback control is performed is required, that means knowledge of the function $\Lambda_N(Y_{N-1})$.

From the integral fluctuation theorem with efficacy parameter (Eq.3.41) it follows a second generalization of the II Law, again from Jensen's inequality:

$$\Phi \geq -\ln \gamma_f. \tag{3.42}$$

The relation between the two fluctuation theorems we derived, one with the measurement information (Eq.3.41), and the other with the efficacy parameter (Eq.3.42), is still under debate. If the fluctuations of the stochastic entropy production $\varphi(X_N|\Lambda_N(Y_{N-1}))$ and stochastic measurement information $I_c(Y_N, X_N)$ are Gaussian distributed, it can be shown that the efficacy parameter γ_f describes the (negative) correlation between $\varphi(X_N|\Lambda_N(Y_{N-1}))$ and $I_c(Y_N, X_N)$, meaning how well on average the obtained information $I_c(Y_n, X_n)$ helps in decreasing the entropy production $\varphi(X_N|\Lambda_N(Y_{N-1}))$ in single realizations of the experiment.

In the case of a gas in contact with a single heat reservoir the entropy production is equivalent to the dissipated work, $\varphi(X_N|\Lambda_N(Y_{N-1})) = \beta\,(W(X_N|\Lambda_N(Y_{N-1})) - \Delta F(\lambda_0, \lambda_N))$. As in the previous section (3.3) we assume the system to start from the canonical distribution at fixed λ_0 and then, at the end of the dynamics, to have time to relax to the canonical equilibrium distribution with λ_N. Then the Sagawa-Ueda integral fluctuation theorem with measurement information (eq.3.38) lead to the generalized Jarzynski equality:

$$\left\langle e^{-W(X_N|\Lambda_N(Y_{N-1}))-I_c^{st}(Y_N,X_N)} \right\rangle = e^{-\Delta F(\lambda_0,\lambda_N)}. \tag{3.43}$$

Szilard engine with measurement errors

The II Law inequality corresponding to Eq.3.43 is:

$$\langle W \rangle \geq \Delta F - K_B T \; I_C \; . \tag{3.44}$$

Let us now consider a generalized Szilard engine with measurements errors[IS11] as a first application of the theory. The control protocol applied on the single-particle gas goes this way: we insert a barrier dividing the volume into two equal parts, and the particle is either in the left half ($x = 0$) or in the right half ($x = 1$) of the volume. We next perform a measurement of the system state with outcome $y = 0$ or $y = 1$, such measurement being characterized by an error rate ϵ: $P(0|0) = P(1|1) = 1 - \epsilon$, $P(1|0) = P(0|1) = \epsilon$. We then move the barrier quasistatically up to a point where

the ratio of the subvolumes is $\frac{v_0}{1-v_0}$ for $y=0$, and $\frac{1-v_1}{v_1}$ for $y=1$ (with $0 \leq v_0 \leq 1$ and $0 \leq v_1 \leq 1$). The last step is the removal of the barrier and the engine returning to its initial state since the previous information is lost. As it is always the case, in a cycle $\Delta F = 0$.

Let us recall that the work in an isothermal expansion between volumes V_f and V_i is given by $W = -K_B T \ln \frac{V_f}{V_i}$. Then the average work extracted $W_{ext} = -W$ in the Szilard engine process is calculated considering the four possible scenarios (x, y) as $\langle W_{ext} \rangle = K_B T \left(\frac{1-\epsilon}{2} \ln(2v_0 v_1) + \frac{\epsilon}{2} \ln(2(1-v_0)(1-v_1)) \right)$. The maximum extracted work for a fixed value of the error rate ϵ is achieved when $v_0 = v_1 = 1-\epsilon$, and its value is $\langle W_{ext} \rangle = (\epsilon \ln(2\epsilon) + (1-\epsilon)\ln(2(1-\epsilon)))$. In this case the measurement information, that is equal to the mutual information in a single measurement process, is given by $I_C = \epsilon \ln(2\epsilon) + (1-\epsilon)\ln(2(1-\epsilon))$. Then the upper bound of extractable work given by Eq.3.44 is achieved by the generalized Szilard engine, $\langle W \rangle = \Delta F - K_B T \; I_C$.

In order to evaluate the efficacy parameter of feedback γ_f (Eq.3.40), we have to consider deterministic backward control protocols $\Lambda_N(Y_{N-1})$ corresponding to measurements on the forward process Y_{N-1}. This amounts to first dividing the box inserting the barrier in order to asymmetrically divide the volume with ratios $\frac{v_0}{1-v_0}$ or $\frac{1-v_1}{v_1}$, corresponding respectively to $y=0$ and $y=1$ in the forward process. Then the barrier is moved quasistatically to the center, measurement is performed with outcome y', and then removed.

Then the efficacy parameter is:

$$\begin{aligned} \gamma_f &= P(y'=0|\Lambda(y=0)) + P(y'=1|\Lambda(y=1)) = \\ &= (1-\epsilon)(v_0+v_1) + \epsilon(2-v_0-v_1) = \left\langle e^{-\beta W} \right\rangle . \end{aligned} \tag{3.45}$$

Here the asymmetry between forward and backward processes is introduced by the fact that measurements occur at different time instants in the forward and backward protocols, and also by the fact that the barrier moves towards the center in the backward process and away from the center in the forward process. In other words, it is the discreteness and time asymmetry of the feedback control scheme itself that imposes the specification of the experiment type. In this setting, three kinds of experiments are needed for the estimation of the entropy production, that is the thermal average of Eq.3.36:
1) Experiment with feedback control.
2) Forward experiment (without feedback control) with fixed control protocol, such protocol resulting from measurements on a dynamics with feedback.
3) Backward experiment (without feedback control) with fixed control protocol,

such protocol being the time-reversal of a protocol resulting from measurements on a dynamics with feedback.

3.5 The Horowitz-Esposito approach

In previous sections (3.4-3.5) we derived classical results of the nonequilibrium thermodynamics of feedback, where a control protocol had to be specified to perform work on the system. In particular we described the Sagawa-Ueda theory of feedback control, and how using information from measurements one can adjust control protocols and extract work from small thermodynamics systems. In the following chapters 4 and 5 we will be mainly interested in the information thermodynamics of autonomous systems, and its application in the reverse engineering of biological systems where we deal with data or data-driven models, and there is no practical possibility of constructing feedback protocols at the fluctuations level.

Here we introduce an alternative formulation of stochastic feedback thermodynamics, where controller and controlled system are described in a symmetric way. This is the the Horowitz-Esposito bipartite network thermodynamics description[HE14; EB10], that we here formalize in terms of bipartite SDEs in the Fokker-Planck representation[BE10]. Here we are specifically interested in their definition of information flow between the two subsystems, and it will introduce our following description of causal influences.

Let us start from the bivariate SDE for the two interacting variables x and y:

$$\begin{cases} dx = g_x(x,y,t)dt + \sqrt{D_x(x,y,t)}\ dW_x \\ dy = g_y(x,y,t)dt + \sqrt{D_y(x,y,t)}\ dW_y \end{cases} \tag{3.46}$$

where $D_x(x,y,t)$ and $D_y(x,y,t)$ are diffusion coefficients, and the (x,y,t) dependence is accounting for the case of multiplicative noise. As usual, Brownian motions are characterized by $\langle dW_i(t)dW_j(t')\rangle = \delta_{ij}\delta_{tt'}dt$, for any $dt > 0$. We note that the system of Eq.3.46 is not a stationary process in general, therefore its probability density depends on the initial distribution $p(x,y,t) \equiv p(x,y,t|p(x,y,t_{in}))$ with $t_{in} \leq t$.

3.5.1 Probability currents and entropy production

This SDE can be transformed into the following Fokker-Planck equation[Ris84] for the time evolution of the probability density $p(x, y, t)$:

$$\begin{aligned}\frac{\partial p(x,y,t)}{\partial t} = &-\frac{\partial}{\partial x}(g_x(x,y,t)p(x,y,t)) - \frac{\partial}{\partial y}(g_y(x,y,t)p(x,y,t)) + \\ &+\tfrac{1}{2}\frac{\partial^2}{\partial x^2}(D_x(x,y,t)p(x,y,t)) + \tfrac{1}{2}\frac{\partial^2}{\partial y^2}(D_y(x,y,t)p(x,y,t)).\end{aligned} \tag{3.47}$$

A rigorous proof of the validity of transformations of stochastic differential equations (like Eq.3.46) into Fokker-Planck equations (like Eq.3.47) for multivariate Markovian systems is given in [Gil96b].

The Fokker-Planck equation (Eq.3.47) can be expressed as a continuity equation in terms of the probability current $\vec{J} = (J_x, J_y)$:

$$\frac{\partial p(x,y,t)}{\partial t} = -\vec{\nabla}\cdot\vec{J}(x,y,t) = -\frac{\partial J_x(x,y,t)}{\partial x} - \frac{\partial J_y(x,y,t)}{\partial y}. \tag{3.48}$$

Comparing Eq.3.48 with Eq.3.47 we identify the currents in the two directions:

$$J_x(x,y,t) = g_x(x,y,t)p(x,y,t) - \frac{1}{2}\frac{\partial}{\partial x}(D_x(x,y,t)p(x,y,t)). \tag{3.49}$$

$$J_y(x,y,t) = g_y(x,y,t)p(x,y,t) - \frac{1}{2}\frac{\partial}{\partial y}(D_y(x,y,t)p(x,y,t)). \tag{3.50}$$

Let us now consider the entropy $S^{xy}(t)$ of the joint system (x, y):

$$S^{xy}(t) = -\int\int dxdx\; p(x,y,t)\ln p(x,y,t). \tag{3.51}$$

The entropy production in the system is the time derivative of the internal entropy $S^{xy}(t)$:

$$\begin{aligned}\tfrac{dS^{xy}(t)}{dt} &= -\textstyle\int\int dxdy\; \frac{\partial p(x,y,t)}{\partial t}\,(1+\ln p(x,y,t)) = -\int\int dxdy\; \frac{\partial p(x,y,t)}{\partial t}\ln p(x,y,t) = \\ &= \textstyle\int\int dxdy\;\vec{\nabla}\cdot\vec{J}(x,y,t)\ln p(x,y,t) = -\int\int dxdy\;\vec{J}(x,y,t)\cdot\frac{\vec{\nabla}p(x,y,t)}{p(x,y,t)} = \\ &= -\textstyle\int\int dxdy\;\frac{1}{p(x,y,t)}\left(J_x(x,y,t)\frac{\partial p(x,y,t)}{\partial x} + J_y(x,y,t)\frac{\partial p(x,y,t)}{\partial y}\right) = \\ = \int\int dxdy\;&\tfrac{1}{p(x,y,t)}\Big[J_x(x,y,t)\tfrac{2}{D_x(x,y,t)}\left(-g_x(x,y,t)p(x,y,t) + J_x(x,y,t) + \tfrac{1}{2}p(x,y,t)\tfrac{\partial D_x(x,y,t)}{\partial x}\right) + \\ &+J_y(x,y,t)\tfrac{2}{D_y(x,y,t)}\left(-g_y(x,y,t)p(x,y,t) + J_y(x,y,t) + \tfrac{1}{2}p(x,y,t)\tfrac{\partial D_y(x,y,t)}{\partial y}\right)\Big],\end{aligned} \tag{3.52}$$

where in the second passage we used the probability normalization $\int\int dxdy\;\frac{\partial p(x,y,t)}{\partial t} = 0$, in the fourth passage we performed partial integration assuming the currents to vanish enough fast for $x, y \to \pm\infty$, and in the last passage we considered the expressions relating the currents and the spatial derivatives (Eq.3.49-3.50).

Following [BE10] we can identify the two terms corresponding to the total (system+reservoir) irreversible entropy production $\dot{S}_i(t) \geq 0$ and the entropy change in the surrounding environment $\dot{S}_r(t)$ due to the interaction with the system, $\dot{S}_i(t) = \frac{dS^{xy}(t)}{dt} + \dot{S}_r(t)$:

$$\dot{S}_r(t) = \int\int dxdy \left[J_x(x,y,t) \frac{2}{D_x(x,y,t)} \left(g_x(x,y,t)p(x,y,t) - \frac{1}{2}p(x,y,t)\frac{\partial D_x(x,y,t)}{\partial x} \right) + \right.$$
$$\left. + J_y(x,y,t) \frac{2}{D_y(x,y,t)} \left(g_y(x,y,t)p(x,y,t) - \frac{1}{2}p(x,y,t)\frac{\partial D_y(x,y,t)}{\partial y} \right) \right]. \tag{3.53}$$

$$\dot{S}_i(t) = \int\int dxdy \, \frac{1}{p(x,y,t)} \left(\frac{2J_x^2(x,y,t)}{D_x(x,y,t)} + \frac{2J_y^2(x,y,t)}{D_y(x,y,t)} \right) \geq 0. \tag{3.54}$$

$\dot{S}_i(t)$ and $\dot{S}_r(t)$ are written as rates because we do not explicitly consider the details of the thermal reservoir. $\dot{S}_r(t)$ quantifies the energy flow to the environment, since the products $g_x(x,y,t)J_x(x,y,t)$ and $g_y(x,y,t)J_y(x,y,t)$ quantify the system energy variations[Che+06; Sei12]. Note that steady-state currents satisfy $\frac{dS^{xy}(t)}{dt} = 0$, and $\dot{S}_i(t) = \dot{S}_r(t)$.

3.5.2 Information flow and thermodynamic inequalities

Horowitz-Esposito[HE14] define "***information flow***" as the time variation of the mutual information $I(t)$ between the two subsystems states:

$$\frac{dI(t)}{dt} = \frac{d}{dt}\int\int dxdy \, p(x,y,t) \ln \frac{p(x,y,t)}{p(x,t)p(y,t)} = \int\int dxdy \, \frac{\partial p(x,y,t)}{\partial t} \ln \frac{p(x,y,t)}{p(x,t)p(y,t)} =$$
$$= \int\int dxdy \, \left(J_x(x,y,t) \frac{\partial \ln p(y,t|x,t)}{\partial x} + J_y(x,y,t) \frac{\partial \ln p(x,t|y,t)}{\partial y} \right) =$$
$$= \dot{I}^x(t) + \dot{I}^y(t), \tag{3.55}$$

where in the third passage we performed partial integration assuming the current to vanish sufficiently fast at infinity. In the last line we defined a decomposition of the information flow into x and y currents:

$$\dot{I}^x(t) = \int\int dxdy \, J_x(x,y,t) \frac{\frac{\partial}{\partial x}p(y,t|x,t)}{p(y,t|x,t)}. \tag{3.56}$$

$$\dot{I}^y(t) = \int\int dxdy \, J_y(x,y,t) \frac{\frac{\partial}{\partial y}p(x,t|y,t)}{p(x,t|y,t)}. \tag{3.57}$$

With this we introduced directionality, and interestingly at steady state ($\frac{dI(t)}{dt} = 0$) the information flow in the Horowitz-Esposito definition[HE14] results to be unidirectional $\dot{I}^x = -\dot{I}^y$. If $I^y > 0$, the dynamics of variable y is creating correlation, meaning that its dynamics is influenced by the position of variable x. In other words $I^y > 0$ means that y is measuring x.

In the same way as we decomposed information flow in x and y currents, we can do the same for the total irreversible entropy production $\dot{S}_i(t) = \dot{S}_i^x(t) + \dot{S}_i^y(t) \geq 0$, for the energy flow to the environment $\dot{S}_r(t) = \dot{S}_r^x(t) + \dot{S}_r^y(t)$, and also for the joint system entropy production $\frac{dS^{xy}(t)}{dt} = (\frac{dS^{xy}(t)}{dt})^x + (\frac{dS^{xy}(t)}{dt})^y$. We see from Eq.3.54 that both terms are positive: $\dot{S}_i^x(t) \geq$, $\dot{S}_i^y(t) \geq 0$. We now want to relate this currents decomposition to the entropy production in each of the two systems alone. This is calculated for system x (can be done for y in the same way) introducing the marginal derivative $\frac{\partial p(x,t)}{\partial t} = \int dy \frac{\partial p(x,y,t)}{\partial t}$ in the time derivative of the entropy $S^x(t)$:

$$\begin{aligned} \frac{dS^x(t)}{dt} &= -\int dx\ \frac{\partial p(x,t)}{\partial t} \ln p(x,t) = \\ = -\int dx \int dy\ \frac{\partial p(x,y,t)}{\partial t} \ln p(x,t) &= -\int\int dxdy\ J_x(x,y,t) \frac{\partial \ln p(x,t)}{\partial x} = \\ &= \dot{I}^x(t) + (\frac{dS^{xy}(t)}{dt})^x = \dot{S}_i^x(t) - \dot{S}_r^x(t) + \dot{I}^x(t), \end{aligned} \tag{3.58}$$

where in the fourth passage we used Eq.3.52 and Eq.3.55. Then using the non-negativity of the irreversible entropy production terms $\dot{S}_i^x(t)$ and $\dot{S}_i^y(t)$ we obtain the **Horowitz-Esposito inequalities**:

$$\begin{cases} \dot{S}_i^x(t) = \frac{dS^x(t)}{dt} + \dot{S}_r^x(t) - \dot{I}^x(t) \geq 0. \\ \dot{S}_i^y(t) = \frac{dS^y(t)}{dt} + \dot{S}_r^y(t) - \dot{I}^y(t) \geq 0. \end{cases} \tag{3.59}$$

The Horowitz-Esposito inequalities describe the stochastic thermodynamics of the interacting subsystems, whose entropy variations are influenced by information flow. This is a further generalization of the Sagawa thermodynamics of feedback control we discussed last paragraph 3.4. The Maxwell's demon here can be identified with one of the two systems, say x, and with an appropriate feedback control to which an information flow is associated he allows for negative values of entropy production when measured on just the controlled system y. In different words, the total entropy production evaluated looking at system y and its interaction with the environment only, can be negative $\frac{dS^y(t)}{dt} + \dot{S}_r^y(t) < 0$ due to the information flow into the measuring system x, $\dot{I}^x(t) > 0$, of course when this information is used to plan an efficient control protocol with $\dot{I}^y(t) < 0$.

3.6 Information flow fluctuations in the feedback cooling model

We introduced stochastic thermodynamics quantities as realization-dependent counterparts of the classical macroscopic heat, work, internal energy, entropy, and information. Indeed, in small systems these quantities can have large fluctuations due to the noise intensity being comparable with the deterministic forces. We also largely discussed how the information from measurement can be used with feedback control

protocols to drive the system entropy production to negative rates thus extracting work. In the last section 3.5 we introduced the Horowitz-Esposito framework for an explicit description of the feedback controller dynamics, meaning that stochastic differential equations can be used for the whole process of measurement and feedback. Following [RH16], we will consider fluctuations of the information flow between system and controller at steady-state in a well studied model of feedback cooling of a Brownian particle[HS14; MR13; MR12]. The particle has constant mass m and variable velocity v_t, and the measurement device position is y_t. The measurement is not instantaneous, but it is a noisy low-pass filter with cut-off frequency $\frac{1}{\tau_f}$. The coupled dynamics is described by the SDE:

$$\begin{cases} m\dot{v} = -\gamma v - ky + \xi_t, \\ \tau_f \dot{y} = v - y + \eta_t, \end{cases} \tag{3.60}$$

where ξ_t and η_t denote Brownian noise sources in the white noise representation. The particle is described by a Langevin equation with friction coefficient γ, where the noise intensity is described at equilibrium (Einstein relation[Phi+12]) by $\langle \xi_t \xi_{t'} \rangle = 2\gamma T \delta(t-t')$, and such relation was postulated to hold also out of equilibrium[Sei05]. The measurement noise is described by $\langle \eta_t \eta_{t'} \rangle = \Delta \delta(t-t')$.

The measurement device y is continuously extracting work from the particle through the feedback term $-ky$ keeping it in a nonequilibrium steady-state with a smaller kinetic temperature compared to the autonomous particle, $T_{kin} \equiv m\langle v^2 \rangle < T$. In other words, y is a refrigerator for x.

The information thermodynamics description in the Horowitz-Esposito framework (introduced in section 3.5) requires to split the probability current in the Fokker-Planck equation into v and y components, $d_t p(v,y,t) = -\partial_v J_v(v,y,t) - \partial_y J_y(v,y,t)$. These are given by (see Eq.3.46-3.47):

$$\begin{cases} J_v(v,y,t) = -\frac{1}{m}(\gamma v + ky)p(v,y,t) - \frac{\gamma T}{m^2}\partial_v p(v,y,t), \\ J_y(v,y,t) = -\frac{1}{\tau_f}(y-v)p(v,y,t) - \frac{\Delta}{2\tau_f^2}\partial_y p(v,y,t). \end{cases} \tag{3.61}$$

We defined the Horowitz-Esposito information flow in Eq.3.56-3.57. Now we consider the stochastic (realization dependent) counterpart of the information flow, and it is related the evolution of the stochastic mutual information between v and y, $I^{st}(t) = \ln \frac{p(v,y,t)}{p(v,t)p(y,t)}$. The information flow is found splitting the total derivative of $I^{st}(t)$ into v and y fluxes: $\frac{dI^{st}(t)}{dt} = i_v(t) + i_y(t)$. The ensemble average of $i_v(t)$ and

$i_y(t)$ is the information flow defined in Eq.3.56-3.57. Let us write the v component of the stochastic information flow:

$$\begin{aligned} i_v(v,y,t) = \tfrac{1}{p(v,y,t)}\partial_v J_v(v,y,t) - \tfrac{1}{p(v,t)}\partial_v J_v(v,t)|_{v(t)} + \\ -\dot{v}\ \partial_v \ln p(v,y,t) + \dot{v}\ \partial_v \ln p(v,t)|_{v(t)}, \end{aligned} \tag{3.62}$$

where $J_v(v,t) = \int dy\ J_v(v,y,t)$ satisfy $d_t p(v,t) = -\partial_v J_v(v,t)$.
$i_v(v,y,t)$ explicitly depends on the instantaneous acceleration $\dot{v}$, that is just a function of the white noise ξ_t when $v(t)$ and $y(t)$ are specified. For the sake of clarity and to uniform the formalism of different authors, we recall that as in previous sections all the quantities including total and partial derivatives are evaluated at (v,y,t) which means $(v(t)=v, y(t)=y, t)$, where the explicit time dependence is taken into account for non-stationary processes. An analogous expression to Eq.3.62 holds for the y component $i_y(v,y,t)$. Let us note that the conditional ensemble average of the acceleration $\dot{v}$ is related to the probability current by definition $\langle \dot{v}|v,y,t\rangle p(v,y,t) = J_v(v,y,t)$. Eq.3.62 describes the dynamics of the stochastic mutual information due to movements of v. The stochastic information flow $i_v(v,y,t)$ has a mixed character since it depends on both the particular realization of the noise ξ_t which determines $\dot{v}$, and on the whole ensemble of trajectories described by $p(v,y,t)$. The idea of considering not just the ensemble averages (or ensemble exponential averages) of stochastic thermodynamics quantities, but also their time evolution in single trajectories dates back to Seifert[Sei05]. He first considered the motion of the stochastic entropy $s(v,y,t) = -\ln p(v,y,t)$, and in our bidimensional problem it can be split again in v and y components. Let us write the v component of the stochastic entropy dynamics:

$$\dot{s}_v(v,y,t) = \frac{\partial_v J_v(v,y,t)}{p(v,y,t)} - \dot{v}\ \partial_v \ln p(v,y,t). \tag{3.63}$$

Rosinberg-Horowitz defined in [RH16] the **time-integrated information current** I^v_{trj} for a trajectory in the time interval $[0,t]$:

$$\begin{aligned} I^v_{trj} \equiv \textstyle\int_0^t dt'\ i_v(v,y,t') = \\ = \textstyle\int_0^t dt'\ \dot{s}_v(v,y,t') + \int_0^t dt'\ \left(\dot{v}(t')\ \partial_v \ln p(v,t') - \frac{\partial_v J_v(v,t')}{p(v,t')}\right) = \\ = \textstyle\int_0^t dt'\ \dot{s}_v(v,y,t') + \int_0^t dt'\ \left(\dot{v}(t')\ \partial_v \ln p(v,t') + \partial_t \ln p(v,t')\right) = \\ = \textstyle\int_0^t dt'\ \dot{s}_v(v,y,t') + \int_0^t dt'\ \frac{d \ln p(v,t')}{dt'} = \\ = \textstyle\int_0^t dt'\ \dot{s}_v(v,y,t') + \ln \frac{p(v,t)}{p(v,0)}. \end{aligned} \tag{3.64}$$

The interaction of the Brownian particle with the thermal reservoir is described by the stochastic entropy production $\varphi^v \equiv \Delta s - \frac{Q}{T}$, that is a sum of the entropy change in the particle $\Delta s = -\ln p(v,t) + \ln p(v,0)$ and of the entropy change in the thermal bath due to the heat exchanged with the particle, $\varphi^v_r = -\frac{Q}{T}$ (see previous

sections 3.3 and 3.4). The latter is written as a Stratonovich integral of the work exchanged with the medium, $Q = \int_0^t dt'(-\gamma v(t') + \xi_{t'}) \circ v(t')$. We agree with this identification of heat in the stochastic energetics of Langevin systems, and it was exhaustively discussed by Sekimoto[Sek10; Sek98]. The choice of the Stratonovich integral is important here, because otherwise the thermal fluctuations described by the Brownian term $\xi_t \sim \frac{dW_t}{dt}$ would have no contribution in the ensemble average of the heat Q.

3.6.1 Modified dynamics, Onsager-Machlup action functionals, and the partial entropy production fluctuation theorem

In the feedback cooling model of Eq.5.43 the macroscopic entropy production $\Phi^v \equiv \langle \varphi^v \rangle < 0$ is negative due to feedback in the stationary cooling regime. An integral fluctuation theorem was derived for the so called "partial entropy production" ϕ^v defined in [RH16] as a sum of the stochastic entropy production φ^v and the integrated information current I^v_{trj}:

$$\phi^v \equiv \varphi^v + I^v_{trj} = \varphi^v_r + \int_0^t dt' \; \dot{s}_v(v, y, t'). \tag{3.65}$$

The fluctuation theorem reads $\langle e^{-\phi^v} \rangle = 1$. Let us denote trajectories in the interval $[0, t]$ as v_0^t and y_0^t, and their backward counterparts as $\widetilde{v_0^t}$ and $\widetilde{y_0^t}$. The particle velocity is odd under time reversal, $\widetilde{v_0^t} = -v_t^0$, while the position is even, $\widetilde{y_0^t} = y_t^0$. If the partial entropy production can be written as $\phi^v = \ln \frac{p(v_0^t, y_0^t)}{p^*(v_0^t, y_0^t)}$, where $p^*(\widetilde{v_0^t}, \widetilde{y_0^t})$ is the probability of observing the backward trajectory in another process, then the fluctuation theorem is proved: $\langle e^{-\phi^v} \rangle = \int\int dv_0^t dy_0^t \; p^*(\widetilde{v_0^t}, \widetilde{y_0^t}) = 1$. Then the derivation consist in finding the particular modified process that gives $\phi^v = \ln \frac{p(v_0^t, y_0^t)}{p^*(\widetilde{v_0^t}, \widetilde{y_0^t})}$. This is obtained[RH16] with a modification in the measurement process only:

$$\tau \dot{y} = v + y + \frac{\Delta}{\tau_f} \partial_y \ln p(-v, y, t) + \eta_t. \tag{3.66}$$

The joint probabilities of trajectories are expressed in terms of Onsager-Machlup action functionals[OM53; MO53; CD01]: $p(v_0^t, y_0^t) = \hat{p}(v_0^t | y_0^t, v(0)) \hat{p}(y_0^t | v_0^t, y(0)) p(v(0), y(0))$. These can be thought as the continuous path-integral limit of Eq.3.33 in the Sagawa-Ueda discrete formalism (discussed in section 3.4). Since the particle dynamics is not changed in the modified process we have $p^*(\widetilde{v_0^t}, \widetilde{y_0^t}) = \hat{p}(\widetilde{v_0^t} | \widetilde{y_0^t}, \widetilde{v(0)}) \hat{p*}(\widetilde{y_0^t} | \widetilde{v_0^t}, \widetilde{y(0)}) p(v(t), y(t))$, where the initial state distribution for the modified dynamics is taken to be the final

state distribution of the standard dynamics. The action functionals are written in the Stratonovich interpretation as:

$$\hat{p}(y_0^t|v_0^t, y(0)) \propto e^{\frac{t}{2\tau_f} - \frac{1}{2\Delta}\int_0^t dt'(\tau_f \dot{y} - v + y)^2} \tag{3.67}$$

$$\hat{p*}(\widetilde{y_0^t}|\widetilde{v_0^t}, \widetilde{y(0)}) \propto e^{-\frac{t}{2\tau_f} - \frac{1}{2\Delta}\int_0^t dt'\left([\tau_f \dot{y} - v + y + \frac{\Delta}{\tau_f}\partial_y \ln p(v,y,t')]^2 + (\frac{\Delta}{\tau_f})^2 \partial_y^2 \ln p(v,y,t')\right)}, \tag{3.68}$$

where we considered that $\dot{y}$ and v change sign in the time-reversal conjugate variables. The detailed fluctuation theorem (Eq.3.27) relates the entropy change in the heat bath to the action functionals, $\varphi_r^v = \ln \frac{\hat{p}(v_0^t|y_0^t, v(0))}{\hat{p}(\widetilde{v_0^t}|\widetilde{y_0^t}, \widetilde{v(0)})}$. Let us now evaluate $\ln \frac{p(v_0^t, y_0^t)}{p^*(\widetilde{v_0^t}, \widetilde{y_0^t})}$ for the modified process (Eq.3.66) and show that is equal to the partial entropy production ϕ^v:

$$\begin{aligned}
\ln \frac{p(v_0^t, y_0^t)}{p^*(\widetilde{v_0^t}, \widetilde{y_0^t})} &= \varphi_r^v + \ln \frac{\hat{p}(y_0^t|v_0^t, y(0))p(v(0), y(0))}{\hat{p*}(\widetilde{y_0^t}|\widetilde{v_0^t}, \widetilde{y(0)})p(v(t), y(t))} = \\
&= \varphi_r^v + \ln \frac{p(v(0), y(0))}{p(v(t), y(t))} + \frac{t}{\tau} + \frac{1}{\tau}\int_0^t dt' [(\tau_f \dot{y} + y - v)\partial_y \ln p(v, y, t') + \frac{\Delta}{2\tau_f}\frac{\partial_y^2 p(v,y,t')}{p(v,y,t')}] = \\
&= \varphi_r^v + \ln \frac{p(v(0), y(0))}{p(v(t), y(t))} - \int_0^t dt'\ \dot{s_y}(v, y, t') = \\
&= \varphi_r^v + \int_0^t dt'\ \dot{s}_v(v, y, t') = \phi^v.
\end{aligned} \tag{3.69}$$

where in the second passage we used $(\partial_y \ln f(y))^2 + \partial_y^2 \ln f(y) = \frac{\partial_y^2 f(y)}{f(y)}$, and in the third passage we used the J_y current expression with $\dot{s}_y(v, y, t) = \frac{\partial_y J_y(v,y,t)}{p(v,y,t)} - \dot{y}\ \partial_y \ln p(v, y, t)$. We then proved the **Rosinberg-Horowitz integral fluctuation theorem** (IFT):

$$\langle e^{-\phi^v} \rangle = \langle e^{-\varphi^v - I_{trj}^v} \rangle = 1. \tag{3.70}$$

The corresponding inequality sets the time-integrated information current as the boundary to the work extracted at steady-state:

$$\langle W_{ext} \rangle = -T \langle \varphi_r^v \rangle \leq T \left\langle I_{trj}^v \right\rangle. \tag{3.71}$$

The v component of the information current enters the inequality, and it quantifies the efficiency of feedback. At steady state $\left\langle I_{trj}^v \right\rangle = -\left\langle I_{trj}^y \right\rangle$. The particular form of the equations in the feedback cooling model did not play a role in the proof of the integral fluctuation theorem (Eq.3.70). Indeed, the IFT for the partial entropy production holds for any coupled Langevin processes involving independent Brownian noise sources[Ros+16]. In a similar way, considering appropriate modified processes, one can find individual IFTs for the dissipated heat, $\langle \frac{p(v(t))}{p(v(0))} e^{-I_{trj}^v} \rangle = e^{\frac{\gamma}{m}t}$, and for the information flow, $\langle e^{-\varphi_r^v} \rangle = e^{\frac{\gamma}{m}t}$.

Eq.3.71 is the second information thermodynamic bound on the extractable work by a Maxwell's demon that we introduce in this chapter. It was shown that the bound based on information flow is more accurate than the transfer entropy bound

of Sagawa-Ueda[Har+16; HS14], $\langle i_v \rangle < T_{x \to y}$. In the next section we will show a recent extension of the Sagawa-Ueda theory made by Sosuke Ito[Ito16] and it shows the equivalence of the two formalisms. For the sake of completeness, we mention that in the refrigerator model an even more accurate bound was provided[MR14; MR13; MR12], and it is the so called "entropy pumping". This is based on a coarse graining of the Fokker-Planck equation with the introduction of the effective feedback force $\bar{f^{fb}}(v,t) \equiv -k\langle y,t|v,t\rangle$. The entropy pumping is defined as $\dot{I_{pump}} \equiv \int dv\ p(v,t)\frac{1}{m}\partial_v \bar{f^{fb}}(v,t)$, it is considered a useful method for linear feedback systems but it is not interpreted as a measure of information, therefore we do not discuss it further.

An exact expression for the average extracted work at steady-state was derived for the feedback-cooling model in [HS14]. The maximum extracted work is achieved for the parameter values $k^{opt} = \gamma(\sqrt{1 + \frac{2T}{\gamma\Delta^2}} - 1)$, and $\tau_f \to 0$, meaning that feedback based on instantaneous measurements is more efficient. In real situations where feedback is necessarily discrete, and the measurement device noise can be time correlated (colored noise), it is better to keep a finite cut-off frequency in the measurement process.

3.7 The II Law-like inequality for non-Markovian dynamics

In previous sections we discussed the stochastic thermodynamics of Markovian (memoryless) systems. The only exception was the measurement process in the Sagawa theory (section 3.4), where the feedback protocol could in principle be constructed taking into account measurements at multiple time instants, but still the measurement process was Markovian. Here we will discuss the generalization to fully non-Markovian bivariate systems as it was recently derived by Sosuke Ito[Ito16]. The dynamics is time discrete, and a path of variable x of lenght l up to time k is denoted $x_k^{(l)} = \{x_k, x_{k-1}, \ \dots\ , x_{k-l+1}\}$. Nevertheless, the dynamics is taken to be *bipartite*: $p(x_{k+1}, y_{k+1}|x_k^{(l)}, y_k^{(l)}) = p(x_{k+1}|x_k^{(l)}, y_k^{(l)}) \cdot p(y_{k+1}|x_k^{(l)}, y_k^{(l)})$. This is an important assumption, and it will lead to a discrepancy with the time series scenario as we will discuss in chapter 5. The non-Markovianity is introduced here as a time delay of n steps in the interactions between variables:

$$\begin{cases} p(x_{k+1}|x_k^{(l)}, y_k^{(l)}) = p(x_{k+1}|x_k, y_{k-n}) \\ p(y_{k+1}|x_k^{(l)}, y_k^{(l)}) = p(y_{k+1}|y_k, x_{k-n}) \end{cases} \tag{3.72}$$

The initialization of the process, that is for $k \leq n$, requires a modification of Eq.3.72 into $p(x_{k+1}|x_k^{(l)}, y_k^{(l)}) = p(x_{k+1}|x_k, y_1)$ and $p(y_{k+1}|x_k^{(l)}, y_k^{(l)}) = p(y_{k+1}|y_k, x_1)$. The

total time length of the process is N. The form of Eq.3.72 can describe, as an example, the effect of a protein that has to be produced in the cytoplasm and then translocated into the nucleus to bind a promoter region, this whole process requiring a non negligible amount of time.

We already defined the stochastic transfer entropy between paths as $T^{st}_{x_k^{(l)}\to y_k^{(l)}} \equiv \ln\left(\frac{P(y_{k+1}^{(l+1)}|x_k^{(l)},y_k^{(l)})}{P(y_{k+1}^{(l+1)}|y_k^{(l)})}\right) = \ln\left(\frac{P(y_{k+1}|x_k^{(l)},y_k^{(l)})}{P(y_{k+1}|y_k^{(l)})}\right)$ (note that $T^{st}_{x_k^{(l)}\to y_k^{(l)}}$ here and in [SU12] corresponds to $T^{st}_{x_k^{(l)}\to y_{k+1}^{(l+1)}}$ in [Ito16]). Backward paths of length l are here defined as $\widetilde{x_k^{(l)}} = \{x_{N-k+1}, x_{N-k+2}, \dots, x_{N-k+l}\} = x^{(l)}_{N-k+l}$, then the *backward transfer entropy* between paths is:

$$T^{st}_{\widetilde{x_k^{(l)}}\to\widetilde{y_k^{(l)}}} \equiv \ln\left(\frac{P(y_{k+1}^{(l+1)}|\widetilde{x_k^{(l)}},\widetilde{y_k^{(l)}})}{P(y_{k+1}^{(l+1)}|\widetilde{y_k^{(l)}})}\right) = \ln\left(\frac{P(\widetilde{y_{k+1}}|\widetilde{x_k^{(l)}},\widetilde{y_k^{(l)}})}{P(\widetilde{y_{k+1}}|\widetilde{y_k^{(l)}})}\right) = \\ = \ln\left(\frac{P(y_{N-k}|x^{(l)}_{N-k+l},y^{(l)}_{N-k+l})}{P(y_{N-k}|y^{(l)}_{N-k+l})}\right). \tag{3.73}$$

The stochastic entropy change Δs_b^x in the thermal bath attached to x should describe exclusively an interaction (or heat exchanged) between subsystem x and the thermal bath, while the influence of y is exerted only on the macroscopic bath entropy production $\langle \Delta s_b^x \rangle$ because it drives the ensemble probability density. Therefore we agree on the definition of Δs_b^x that is given in [Ito16]: $\Delta s_b^x \equiv \sum_{k=1}^{N-1} \ln \frac{p(x_{k+1}|x_k^{(l)},y_k^{(l)})}{p_B(x_k|x_{k+1},x_{k-1}^{(l-1)},y_k^{(l)})}$. For the specific process of Eq.3.72 it reads:

$$\Delta s_b^x \equiv \sum_{k=1}^{N-1} \ln \frac{p(x_{k+1}|x_k, y_{k-n})}{p_B(x_k|x_{k+1}, y_{k-n})}, \tag{3.74}$$

where the same condition on the dynamics (y_{k-n}) is imposed for both forward and backward probabilities. The backward probability $p_B(x_k|x_{k+1}, y_{k-n})$ is defined as the conditional probability of observing x_k at time $k+1$ given x_{k+1} at time k and y_{k-n} at time $k-n$. The total stochastic entropy production of x and the thermal bath attached to x is written $\Delta s^x_{tot} = \Delta s_x + \Delta s^x_B$, where the entropy change in the system x during the whole process is $\Delta s_x = \ln \frac{p(x_1)}{p(x_N)}$. As usual, the thermal averages are denoted with $\Delta S^x_{tot} \equiv \langle \Delta s^x_{tot} \rangle$.

Let us now define the directed information[Mas90] $I(x_N^{(N)} \to y_N^{(N)})$ from x paths to y paths as:

$$I(x_N^{(N)} \to y_N^{(N)}) \equiv I(x_1, y_1) + \sum_{k=1}^{n} T_{x_1^{(1)}\to y_k^{(k)}} + \sum_{k=n+1}^{N-1} T_{x_{k-n}^{(1)}\to y_k^{(k)}}. \tag{3.75}$$

Similarly one can define the backward directed information as:

$$I(\widetilde{x_N^{(N)}} \to \widetilde{y_N^{(N)}}) \equiv I(x_N, y_N) + \sum_{k=1}^{n} T_{\widetilde{x_1^{(1)}} \to \widetilde{y_k^{(k)}}} + \sum_{k=n+1}^{N-1} T_{x_{k-n}^{(1)} \to \widetilde{y_k^{(k)}}}. \tag{3.76}$$

Developing the terms one recognizes that:

$$\Delta S_{tot}^{x} + I(x_N^{(N)} \to y_N^{(N)}) - I(\widetilde{x_N^{(N)}} \to \widetilde{y_N^{(N)}}) = \sum_{x_N^{(N)}, y_N^{(N)}} p(x_N^{(N)}, y_N^{(N)}) \ln \frac{p(x_N^{(N)}, y_N^{(N)})}{\widetilde{p}(x_N^{(N)}, y_N^{(N)})} \geq 0, \tag{3.77}$$

where the backward probability $\widetilde{p}(x_N^{(N)}, y_N^{(N)})$ is defined as:

$$\begin{aligned} \widetilde{p}(x_N^{(N)}, y_N^{(N)}) \equiv p(x_N, y_N) \prod_{k=1}^{n} p_B(x_k | x_{k+1}, y_1) \prod_{m'=N-n}^{N-1} p(y_{m'} | y_{m'+1}, x_N) * \\ * \prod_{m=1}^{N-n-1} p(y_m | y_{m+1}, x_{m+n+1}) \prod_{k'=n+1}^{N-1} p_B(x_{k'} | x_{k'+1}, y_{k'-n}). \end{aligned} \tag{3.78}$$

$\widetilde{p}(x_N^{(N)}, y_N^{(N)})$ is different from the probability of backward paths $\widetilde{p}(x_N^{(N)}, y_N^{(N)}) \neq p(\widetilde{x_N^{(N)}}, \widetilde{y_N^{(N)}})$, but it is easily seen to be normalized as well. The **Thermodynamics II Law-like inequality for non-Markovian dynamics** is then written:

$$\begin{aligned} -\Delta S_{tot}^{x} \leq I(x_1, y_1) - I(x_N, y_N) + \sum_{k=1}^{n} \left(T_{x_1^{(1)} \to y_k^{(k)}} - T_{\widetilde{x_1^{(1)}} \to \widetilde{y_k^{(k)}}} \right) + \\ + \sum_{k=n+1}^{N-1} \left(T_{x_{k-n}^{(1)} \to y_k^{(k)}} - T_{x_{k-n}^{(1)} \to \widetilde{y_k^{(k)}}} \right). \end{aligned} \tag{3.79}$$

Note that this result incorporates the Markovian case $n = 0$ where the first sum in the RHS vanishes. For the bivariate Langevin system of section 3.6 the inequality of Eq.3.79 is equivalent to the one of Rosinberg-Horowitz with information flow (Eq.3.71).

4 Causal influence

> *"The causes of all the appearances in nature are the conditions under which they reliably emerge".*
>
> **— Arthur Schopenhauer**

This chapter is meant to introduce the reader to our definition of causal influence, that is published in **Physical Review E 95, 042315** [Auc+17], and to comment the results of that paper. The material in here summarizes, completes and fill the gaps in the paper, so that the reader can have a clear idea of what has been done there. Nevertheless we encourage the reading of our original work [Auc+17], because the flow of ideas and the sentences used there to motivate the concepts are selected with the highest accuracy. Some of those sentences are equivalently repeated here, especially for the enumeration of mathematical properties in the second part where no other rephrasing was possible.

4.1 Introduction to the quantitative definition of causal influence

The concept of causation between observable events is fundamental in the formalization and communication of scientific results, but it still remained rather vague and unprecise in its definition. Causal relations manifest as predictive information and can be inferred from observations, but the quantification of such causal information is not a simple task.

The reader might wonder why we need to talk about a "causal influence" when interactions between observables are symmetric in any microscopic physical theory. That is true, such concept does not apply to any fundamental physics description, and this is also for another reason: in classical physics the dynamics is deterministic, and the absence of uncertainty prevents the definition of information measures. It is clear that the concept of causation is *practical*, and it is appropriate only in coarse grained macroscopic descriptions. Indeed, all the differential equation models in systems biology[Kli+16] present nonphysical asymmetric interactions, and this

is also the case in models from geophysics, financial markets, social behavior, or to put it short *complex systems*. In addition, the stochasticity is there added as a form of uncertainty in the dynamics that derives from a lack of knowledge in the observation of a (supposedly) deterministic underlying dynamics. Let us say that in general we accept a stochastic dynamical description of complex systems with asymmetric interactions. Then the concept of causal influence describes the effect of such asymmetric interactions quantifying the amount of directed non-redundant information flow over time, as we will explain in the following, and already introduced in Chapter 1.

The main intuition is that correlation (or the mutual information) has something to do with causation but is still different from that, because causation is clearly an asymmetric relation while the mutual information is symmetric between variables. Then one asks for a temporal order between events to associate directionality to interactions, thus determining a cause-effect relationship. This can be done looking at the asymmetry of time-lagged correlations, or better at transfer entropy measures. The transfer entropy[Sch00], a generalization of the Granger causality to consider also nonlinear effects[Bar+09], is a widely recognized measure of directed information flow. It is a key quantity in stochastic thermodynamics[Par+15] as we discussed in Chapter 3, and it is widely used in data analysis[Din+06].

Recently the transfer entropy has been criticized as a measure of information flow because it is based on conditioning, and is therefore dominated by synergistic effects[Jam+16]. Then people started reconsidering the problem of defining a measure of information flow between observables. In general these observables can be dynamical systems trajectories, or experimental data in the form of time series. Importantly, the concept of causation is well defined only with respect to a chronology[Sch12].

In the case of continuous trajectories with bipartite structure it seems reasonable to adapt the Horowitz-Esposito information flow[HE14; HS14] we introduced in section 3.5. Let us recall the definition of bipartite dynamics, that is proper of all the models we considered so far and especially in chapter 3. Indeed all the information thermodynamics theory we introduced is valid only for bipartite (or multipartite) dynamics. Bipartite dynamics of two variables x and y is defined by the updating property $p(x_{t+dt}, y_{t+dt}|x_t, y_t) = p(x_{t+dt}|x_t, y_t) \cdot p(y_{t+dt}|x_t, y_t)$. Now the bipartite (or multipartite) structure is very specific, and is almost never found in time series data[Auc+19b; Auc+18]. Indeed in bivariate time series with observational time τ the evolution joint probability is $p(x_{t+\tau}, y_{t+\tau}|x_t, y_t) = p(x_{t+\tau}|x_t, y_t) \cdot p(y_{t+\tau}|x_t, y_t, x_{t+\tau})$, and it becomes bipartite only in the (uninteresting) case of no interaction.

A general definition of directed information flow between stochastic dynamical variables x and y, that is the causal influence, will be defined as a function of the probability of time series realizations. We find the partial information decomposition[WB10; Bar15] a potentially really powerful approach. It consist in the identification of the synergistic contribution to the transfer entropy, or equivalently in the identification of the redundant information contribution to the time-lagged mutual information. We will describe this more in depth in the following sections. Let us just say that, once a partial information decomposition strategy has been selected, one ends up with a so-called "unique information", that should describe the non-redundant information flow. The unique information is in our opinion the right approach for defining the causal influence in a time series setting. In particular, consider the Markovian stationary bivariate stochastic process (x, y), and let its time series be described by the joint probability density $p(x_t, y_t, x_{t+\tau}, y_{t+\tau})$, where τ is the observational time of the process. Then the unique information, or the causal influence $C_{x\to y}(\tau)$ from x to y, should quantify that part of the time-lagged mutual information $I(x_t, y_{t+\tau})$ between x and the evolution of y at time $t+\tau$, that is not already know from y at time t:

$$C_{x\to y}(\tau) = I(x_t, y_{t+\tau}) - R(x_t, y_t; y_{t+\tau}), \tag{4.1}$$

where $R(x_t, y_t; y_{t+\tau})$ is the redundancy measure. The causal influence $C_{x\to y}(\tau)$ quantifies how x is influencing the dynamics of y with the unique (non redundant) information that it gives on its evolution.

Now the difficult part is the definition of the redundancy $R(x_t, y_t; y_{t+\tau})$ (or equivalently the definition of a synergy measure), and many proposals were already there[Auc+17; WB10; GK14; Har+13; Gri+14; Ber+14]. All these definitions we found in the literature, were all demonstrated to take a trivial form in Gaussian systems[Bar15] as the minimum between the information on $y_{t+\tau}$ given by x_t and y_t: $R(x_t, y_t; y_{t+\tau}) = \min[I(x_t, y_{t+\tau}), I(y_t, y_{t+\tau})]$. Importantly, this definition is independent of the mutual information between the sources x_t and y_t, that is $I_{xy} \equiv I(x_t, y_t)$.

This motivated our search for a new definition, sensitive to I_{xy}. Introducing the total predictive information that the two sources give on the target $I_{tot} \equiv I(y_{t+\tau}, (x_t, y_t))$, our definition of redundant information is:

$$R(\tau) \equiv \frac{1}{2} \ln\left(\frac{e^{2(I_{xy}+I_{tot})}}{e^{2I_{xy}} + e^{2I_{tot}} - 1} \right). \tag{4.2}$$

The paper [Auc+17] is basically a motivation and discussion of our definition of redundancy (4.2) and causal influence (4.1) for the case of linear Langevin networks without feedbacks. The main resulting properties of the causal influence $C_{x\to y}(\tau)$

in linear Langevin networks without feedback are that it is zero for $\tau = 0$, it is continuous and positive for $\tau > 0$ meaning that the macroscopic effects of the interaction are seen gradually over time, and it reaches a peak for a finite τ differently from the transfer entropy that can diverge in the limit $\tau \to 0$ because of synergistic contributions, as we will discuss. $C_{x\to y}(\tau)$ vanishes for $\tau \to \infty$ because the effects of past interactions relax over time in stationary systems. Another important feature that is shared with the transfer entropy but not with the mutual information or correlations, is that the causal influence is zero in the absence of direct (or mediated) interaction. In particular we showed how a third object creating time-delayed correlation between two variables x and y results in zero causal influence.

Let us just mention that some scientists following Judea Pearl[Pea95; Pea09] argued that a definition of causal influence is generally not possible in terms of information from observations because of confounding factors, and the only way to infer causality is the possibility to directly manipulate systems and observe responses, with the so called do-calculus. We do not fully agree with this view and still consider causality as a form of unbiased information flow and predictability improvement based on observations[Sch12]. Furthermore, there are many examples where a perturbation of the system is not even possible. As an example, let us say that we wish to quantify the causal influence in the interaction between prices of two stocks in the market. Then it is just not possible for a normal person to buy a sufficiently large number of shares to study the response of the market, unless this person is very rich and the asset without a strong financial liquidity[Shr04].

Our measure of causal influence was inspired by the information processing properties of the basic linear response model (BLRM), that is just discussed shortly in the paper. The BLRM is the simplest of signal-response models, those characterized by the absence of feedback. In the bivariate case (x, y) it allows the identification of an input (signal) and an output (response). A detailed introduction and discussion of the BLRM is provided here with analytical results for the time-lagged mutual information and transfer entropy whose simple derivation we omitted in the paper. We will discuss the causal influence measure in the BLRM first, and then in a three dimensional system with signal-response structure.

In general, the causal influence quantifies the effective strength of asymmetric causal interactions and the time scale over which the effects are seen. Our measure is a good description of the dynamics of influences in linear response models. We will discuss the difficulties in generalizing the causal influence measure to nonlinear and feedback systems.

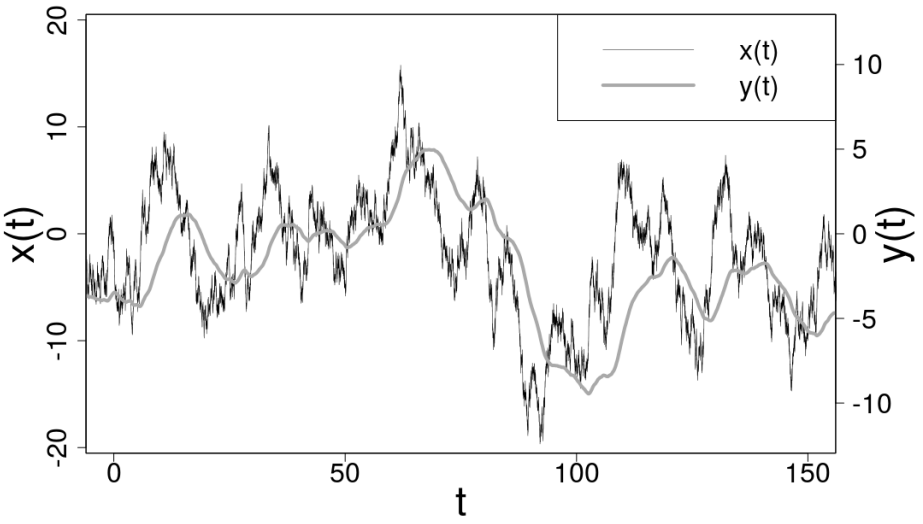


Fig. 4.1: Stochastic dynamics of the Basic Linear Response Model. The parameters are $\alpha = 0.1$, $\beta = 0.2$, $t_{rel} = 10$ and $D = 10$. Figure taken from [Auc+17].

4.1.1 The basic linear response model

The basic linear response model (BLRM) is composed of a fluctuating signal x described by the Ornstein-Uhlenbeck (OU) process[UO30; Gil96a], and a dynamic linear response y to this signal:

$$\begin{cases} dx = -\frac{x}{t_{rel}}dt + \sqrt{D}\ dW \\ \frac{dy}{dt} = \alpha x - \beta y \end{cases} \tag{4.3}$$

where dW is Brownian noise described by $\langle dW_t dW_{t'} \rangle = \delta_{tt'} dt$. A sample of the dynamics is plotted in Fig.4.1. Note that the OU process in (4.3) is written in the white noise representation (see section2.2.1) in the paper, and its solution for the evolution conditional probability we already discussed in section 2.3, and can be summarized with:

$$P(x_{t+t'}|x_t) = \mathcal{N}\left[x_t e^{-\frac{|t'|}{t_{rel}}}, \sigma_x^2(1 - e^{-\frac{2|t'|}{t_{rel}}})\right], \tag{4.4}$$

where $\sigma_x^2 = D\frac{t_{rel}}{2}$. Here we adopt the Ito calculus, but it will lead to no difference in the results because we excluded multiplicative noise and diffusion coefficients are constants (see section 2.5). In particular we write x as a function of the Brownian motion realization with:

$$x_t = \sqrt{D}\int_{-\infty}^{t} dW'_t e^{-\frac{t-t'}{t_{rel}}} \tag{4.5}$$

Then the formal solution of the y dynamics as a function of the noise realization is:

$$y_{t+\tau} = y_t e^{-\beta\tau} + \alpha \int_0^\tau dt'\, x_{t+t'} e^{-\beta(\tau-t')} =$$
$$= y_t e^{-\beta\tau} + \alpha \int_0^\tau dt'\, e^{-\beta(\tau-t')} \int_{-\infty}^{t+t'} \sqrt{D} e^{-\frac{t+t'-t''}{t_{rel}}} dW_{t''}, \tag{4.6}$$

where the last integral has to be interpreted in the Ito sense, meaning that Brownian increments are independent from the variables at the same or previous time points, $\langle x_t dW_t \rangle = \langle x_t \rangle \langle dW_t \rangle = 0$. The BLRM (4.3) is a stationary stochastic process, therefore thermal averages like $\langle y_t^2 \rangle$ or $\langle x_t y_{t+\tau} \rangle$ are independent of the specific time point t, and depend only on the time lag τ. Then we can calculate those by imposing to 0 the derivative with respect to t:

$$0 = \frac{d}{dt}\langle y_t^2 \rangle = 2\langle y_t \frac{dy_t}{dt} \rangle = 2\alpha \langle x_t y_t \rangle - 2\beta \langle y_t^2 \rangle. \tag{4.7}$$

Then we have the relation $\langle y_t^2 \rangle = \frac{\alpha}{\beta}\langle x_t y_t \rangle$, and we proceed to calculate $\langle x_t y_t \rangle$:

$$0 = \tfrac{d}{dt}\langle x_t y_t \rangle = \langle \tfrac{dx_t}{dt} y_t \rangle + \langle x_t \tfrac{dy_t}{dt} \rangle = -\tfrac{1}{t_{rel}}\langle x_t y_t \rangle + \sqrt{D}\tfrac{1}{dt}\langle dW_t y_t \rangle + \alpha \langle x_t^2 \rangle - \beta \langle x_t y_t \rangle =$$
$$= -(\beta + \tfrac{1}{t_{rel}})\langle x_t y_t \rangle + \alpha \sigma_x^2, \tag{4.8}$$

where we used $\sigma_x^2 = \langle x_t^2 \rangle = D\frac{t_{rel}}{2}$ since $\langle x_t \rangle = 0$ as discussed in section 2.3, and the noise property of Ito calculus $\langle dW_t y_t \rangle = 0$. Note that also for y it holds $\langle y_t \rangle = 0$, and $\sigma_y^2 = \langle y_t^2 \rangle$. Then we obtain $\langle x_t y_t \rangle = \frac{\alpha \sigma_x^2}{\beta + \frac{1}{t_{rel}}}$, and $\sigma_y^2 = \frac{\alpha^2 \sigma_x^2}{\beta(\beta + \frac{1}{t_{rel}})}$. Let us now proceed with the calculation of the time-lagged correlation $C(x_t, y_{t+\tau}) = \frac{\langle x_t y_{t+\tau} \rangle - \langle x_t \rangle \langle y_{t+\tau} \rangle}{\sigma_x \sigma_y} = \frac{\langle x_t y_{t+\tau} \rangle}{\sigma_x \sigma_y}$:

$$0 = \tfrac{d}{dt}\langle x_t y_{t+\tau} \rangle = \langle \tfrac{dx_t}{dt} y_{t+\tau} \rangle + \langle x_t \tfrac{dy_{t+\tau}}{dt} \rangle =$$
$$= -(\beta + \tfrac{1}{t_{rel}})\langle x_t y_{t+\tau} \rangle + \alpha \langle x_t x_{t+\tau} \rangle + \alpha D \int_0^\tau dt' e^{-\beta(\tau-t')} \int_{-\infty}^{t+t'} \tfrac{1}{dt}\langle dW_t dW_{t''} \rangle e^{-\frac{t+t'-t''}{t_{rel}}} =$$
$$= -(\beta + \tfrac{1}{t_{rel}})\langle x_t y_{t+\tau} \rangle + \alpha \sigma_x^2 e^{-\frac{\tau}{t_{rel}}} + \alpha D \frac{e^{-\frac{\tau}{t_{rel}}} - e^{-\beta\tau}}{\beta - \frac{1}{t_{rel}}}, \tag{4.9}$$

where we used the noise property $\langle dW_t dW_{t''} \rangle = \delta_{tt''} dt$. Note that the Kronecker delta $\delta_{tt''} dt$ in Ito calculus takes the role of the Dirac delta in Stratonovich calculus, and the two expressions differ in the case of multiplicative noise (see section 2.5). The signal autocorrelation was calculated as $\langle x_t x_{t+\tau} \rangle = \int\int dx_t\; p(x_t) \langle x_{t+\tau} | x_t \rangle x_t = \sigma_x^2 e^{-\frac{\tau}{t_{rel}}}$. Rearranging the terms in (4.9) we obtain the time-lagged correlation:

$$\langle x_t y_{t+\tau} \rangle = \frac{2\alpha t_{rel} \sigma_x^2}{\beta^2 t_{rel}^2 - 1} \left(\frac{\beta t_{rel} + 1}{2} e^{-\frac{\tau}{t_{rel}}} - e^{-\beta\tau} \right). \tag{4.10}$$

Now we are interested in the response time of the BLRM in the statistical sense, that is the time lag τ_{opt} that corresponds to the maximal mutual information, and it

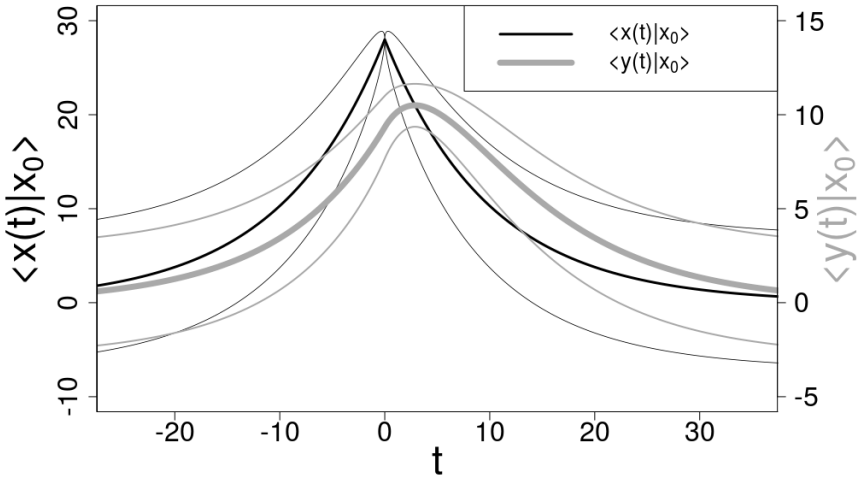


Fig. 4.2: Conditional probability distributions over time. Given a particular condition (input) at time $t = 0$, $x(0) \equiv x_0 = 28$, we plot the conditioned expectation values $\langle y(t)|x(0)\rangle$, $\langle x(t)|x(0)\rangle$ with the relative standard deviations $\pm\sigma_{y(t)|x(0)}$, $\pm\sigma_{x(t)|x(0)}$ (thinner lines) as a function of the time shift t. The parameters are $\alpha = 0.1$, $\beta = 0.2$, $t_{rel} = 10$, $D = 10$. The plot is taken from [Auc+17].

is found imposing $0 = \partial_\tau \langle x_t y_{t+\tau}\rangle |_{\tau_{opt}}$. Then one obtains $\tau_{opt} = \frac{t_{rel}}{\beta t_{rel}-1} \ln\left(\frac{2\beta t_{rel}}{\beta t_{rel}+1}\right)$, that is equation (11) in the paper. The dynamic nature of the response creates the positive lag τ_{opt}. This can even be beneficial in terms of information transmission in (nonlinear) switch-like systems as shown by Nemenmann[Nem12], and not discussed here. The optimal correlation is calculated as:

$$C(x_t, y_{t+\tau_{opt}}) = \frac{\langle x_t y_{t+\tau_{opt}}\rangle}{\sigma_x \sigma_y} = \sqrt{2}\left(\frac{\beta t_{rel}+1}{2\beta t_{rel}}\right)^{\frac{\beta t_{rel}+1}{2(\beta t_{rel}-1)}}, \tag{4.11}$$

which importantly depends on the parameters only through the product βt_{rel}, that is the ratio of the two time scales of the BLRM: signal fluctuations relaxation time t_{rel} and deterministic response time $\frac{1}{\beta}$. In Gaussian systems the mutual information is related to the correlation[CT12] through $I(x_t, y_{t+\tau}) = -\frac{1}{2}\ln(1 - C^2(x_t, y_{t+\tau}))$. Applying this formula to (4.11) one obtains Eq.12 in the paper. The optimal mutual information I^{opt} can be considered as a form of dynamical channel capacity.

In Fig.4.2 are represented the conditional probability densities $p(y_{t+\tau}|x_t)$ and $p(x_{t+\tau}|x_t)$. These are Gaussian distributed, so that only conditional expectations and variances need to be specified. The time lag of the peak of the conditional mean $\langle y(t)|x(0)\rangle$ is due to the fact that the response is dynamic and not immediate, it is a low-pass filter integrating the signal.

Let us now calculate the conditional mutual information or transfer entropy $T_{x\to y}(\tau) \equiv I(x_t, y_{t+\tau}|y_t)$. The chain rule for the mutual information[CT12] reads:

$$I\left((x_t, y_t), y_{t+\tau}\right) = I(y_t, y_{t+\tau}) + I(x_t, y_{t+\tau}|y_t). \tag{4.12}$$

Then in order to calculate the transfer entropy we will first consider the mutual information $I(y_t, y_{t+\tau})$ and the total (mutual) information $I_{tot} \equiv I\left((x_t, y_t), y_{t+\tau}\right)$.

Let us make explicit the no-feedback property of the BLRM and of signal-response models in general. That is specified by the fact that y_t and $x_{t+\tau}$ are conditionally independent given x_t, in formulae $p(y_t, x_{t+\tau}|x_t) = p(y_t|x_t) \cdot p(x_{t+\tau}|x_t)$. Then the correlation $\langle y_t x_{t+\tau}\rangle$ can be calculated with the Chapman-Kolmogorov formula[Gar09; VK92] $p(y_t|x_{t+\tau}) = \int dx_t\ p(y_t|x_t)p(x_t|x_{t+\tau})$:

$$\langle y_t x_{t+\tau}\rangle = \int dx_t\ p(x_t)\langle y_t x_{t+\tau}|x_t\rangle = \int dx_t\ p(x_t)\langle y_t|x_t\rangle\langle x_{t+\tau}|x_t\rangle = \frac{\sigma_x^2 \alpha t_{rel}}{\beta t_{rel}+1} e^{-\frac{\tau}{t_{rel}}}, \tag{4.13}$$

where we used $\langle y_t|x_t\rangle = \alpha \int_{-\infty}^{t} dt'\ \langle x_{t'}|x_t\rangle e^{-\beta(t-t')} = x_t \frac{\alpha t_{rel}}{\beta t_{rel}+1}$. Then we proceed to calculate the correlation $\langle y_t y_{t+\tau}\rangle$=

$$0 = \frac{d}{dt}\langle y_t y_{t+\tau}\rangle = \alpha\langle x_t y_{t+\tau}\rangle - 2\beta\langle y_t y_{t+\tau}\rangle + \alpha\langle y_t x_{t+\tau}\rangle, \tag{4.14}$$

so that we obtain the response autocorrelation:

$$C(y_t, y_{t+\tau}) = \frac{\langle y_t y_{t+\tau}\rangle}{\sigma_y^2} = \frac{e^{-\frac{\tau}{t_{rel}}}\beta t_{rel} - e^{-\beta\tau}}{\beta t_{rel} - 1}. \tag{4.15}$$

The response mutual information is then given by $I(y_t, y_{t+\tau}) = -\frac{1}{2}\ln\left(1 - C^2(y_t, y_{t+\tau})\right)$.

The BLRM is linear and the joint probabilities of all variables are Gaussian, therefore we can calculate the total predictive information as $I_{tot} \equiv I\left((x_t, y_t), y_{t+\tau}\right) = \ln \frac{\sigma_y}{\sigma_{y_{t+\tau}|x_t,y_t}}$, where we introduced the variance $\sigma^2_{y_{t+\tau}|x_t,y_t} = \left\langle (y_{t+\tau} - \langle y_{t+\tau}|x_t, y_t\rangle)^2\right\rangle$. When x_t and y_t are known, then the uncertainty on $y_{t+\tau}$ is only due to the noise realization in the interval $[t, t+\tau)$:

$$y_{t+\tau} - \langle y_{t+\tau}|x_x, y_t\rangle = \alpha\sqrt{D}\int_0^\tau dt'\ e^{-\beta(\tau-t')}\int_0^{t'} dW_{t+t''} e^{-\frac{t'-t''}{t_{rel}}}, \tag{4.16}$$

which is the same for any condition (x_t, y_t) as a consequence of the BLRM linearity. Note that $y_{t+\tau}$ is not generic in (4.16), but is sampled from the conditional $p(y_{t+\tau}|x_t, y_t)$. Then we have:

$$\sigma^2_{y_{t+\tau}|x_t,y_t} = \left\langle (y_{t+\tau} - \langle y_{t+\tau}|x_t, y_t\rangle)^2 \right\rangle =$$
$$= \alpha^2 D e^{-2\beta\tau} \int_0^\tau \int_0^\tau dt' dt'' \; e^{(\beta - \frac{1}{t_{rel}})(t'+t'')} \int_0^{t'} \int_0^{t''} \langle dW_{t+t'''} dW_{t+t''''} \rangle e^{\frac{t'''+t''''}{t_{rel}}} =$$
$$= \alpha^2 D e^{-2\beta\tau} 2 \int_0^\tau \int_0^{t'} dt' dt'' \; e^{(\beta - \frac{1}{t_{rel}})(t'+t'')} \int_0^{t''} dt''' e^{\frac{2t'''}{t_{rel}}} =$$
$$= \frac{\sigma_y^2}{(\beta t_{rel}-1)^2} \big[(\beta t_{rel} - 1)^2 - e^{-\frac{2\tau}{t_{rel}}} \beta t_{rel}(\beta t_{rel} + 1) +$$
$$+ 4\beta t_{rel} e^{-(\beta + \frac{1}{t_{rel}})\tau} - e^{-2\beta\tau}(\beta t_{rel} + 1) \big], \tag{4.17}$$

where we used the integrand symmetry and integrated only in the region $t'' < t'$ (and correspondingly $t'''' < t'''$), and the response variance $\sigma_y^2 = \frac{\alpha^2 \sigma_x^2}{\beta(\beta + \frac{1}{t_{rel}})}$. Then we get the total predictive information $I_{tot} \equiv I\left((x_t, y_t), y_{t+\tau}\right) = \ln \frac{\sigma_y}{\sigma_{y_{t+\tau}|x_t,y_t}}$, and using the chain rule (4.12) we get the transfer entropy:

$$T_{x\to y}(\tau) \equiv I(x_t, y_{t+\tau}|y_t) = I\left((x_t, y_t), y_{t+\tau}\right) - I(y_t, y_{t+\tau}) =$$
$$= \frac{1}{2} \ln \left(1 + \frac{\beta t_{rel} (e^{-\frac{\tau}{t_{rel}}} - e^{-\beta\tau})^2}{(1-\beta t_{rel})^2 - e^{-2\beta\tau}(1+\beta t_{rel}) + e^{-(\beta + \frac{1}{t_{rel}})\tau} 4\beta t_{rel} - e^{-\frac{2\tau}{t_{rel}}} \beta t_{rel}(1+\beta t_{rel})} \right), \tag{4.18}$$

that is Eq.13 in the paper. Note that we corrected a printing error in [Auc+17]: on the numerator it was t_c instead of t_{rel}. A similar relation I already obtained for biochemical fluctuations of receptor-ligand systems during a short collaboration period with Juergen Pahle at the BioQuant in Heidelberg in August 2015. Indeed, receptor-ligand systems in the linear regime are modeled by a BLRM where α and β are replaced by reaction kinetic parameters. We will consider receptor-ligand systems with a more general (nonlinear) model in chapter 5, when discussing the time series thermodynamics of signal-response models.

From the quantities obtained we easily derive formulae (9)-(10) in the paper, that is the relation between the mutual information and the signal-to-noise ratio:

$$I(x(t), y(t+t')) = \frac{1}{2} \ln(1 + \mathbf{SNR}), \tag{4.19}$$

$$\mathbf{SNR} = \frac{(\frac{\partial \langle y(t+t')|x(t)\rangle}{\partial x(t)})^2 \sigma_x^2}{\sigma^2_{y(t+t')|x(t)}}. \tag{4.20}$$

This is a form of the fluctuation-dissipation theorem in linear response theory[Mar+08; Kub66; Kub57].

We calculated the transfer entropy[Sch00] considering the chain rule for the mutual information (4.12). We could as well consider that the transfer entropy is equivalent

to the Granger Causality[Gra69] in Gaussian systems[Bar+09]. Let us recall the definition of Granger causality, $T^G_{x\to y}(\tau) \equiv \left\langle \ln\left(\frac{\sigma_{y_{t+\tau}|y_t}}{\sigma_{y_{t+\tau}|x_t,y_t}}\right)\right\rangle$.

It results that $T_{y\to x}(\tau) = 0$ (recall that by definition $\tau \geq 0$) as it should be since the signal dynamics is independent of the response (no feedback). $T_{x\to y}(\tau)$ is always positive instead and diverges for $\tau \to 0$. This is due to the synergistic interaction between the knowledge of $x(t)$ and $y(t)$ that manifest in the prediction of $y(t+\tau)$. Note that for small τ the variation $y_{t+\tau} - y_t$ is uncertain with order $\tau^{\frac{3}{2}}$, $y(t+\tau) - y(t) = \tau(\alpha x(t) - \beta y(t)) + B(\tau^{\frac{3}{2}})$, while with the knowledge of only y_t the uncertainty is of order τ. This $B(\tau^{\frac{3}{2}})$ is different from the wrong $B(\tau^2)$ we claimed in [Auc+17], and here we corrected it. Note that the uncertainty on $x_{t+\tau}$ is of order $\sqrt{\tau}$ for small τ as it is always the case for Brownian increments.

4.2 Information decomposition and causal influence

We would like to exclude the synergistic effect we just discussed in the last section from the causal influence, that is instead the macroscopic effect of the signal-response asymmetric interaction that is obtained gradually over time after the instantaneous "cause" $x(t)$. It is the information that the signal $x(t)$ gives on the evolution of the response $y(t+\tau)$ that is not redundant in the knowledge of the response $y(t)$.

We adopt a **partial information decomposition** scheme[WB10; Bar15] (PID) of the total information that $x(t)$ and $y(t)$ give on the evolution of the response $y(t+\tau)$:

$$I(y(t+\tau), (x(t), y(t))) = R + U_x + U_y + S, \tag{4.21}$$

where R is the redundancy, U_x and U_y are the unique information contributions respectively of $x(t)$ and $y(t)$ alone, and the synergy S is defined as the information that one gets in addition when considering simultaneously both $x(t)$ and $y(t)$. Note that $I(x(t), y(t+\tau)) = R + U_x$ and $T_{x\to y} = U_x + S$.

The challenge in the community[Jam+16] is a definition of redundancy that gives the unique informations U_x and U_y the form of a (discrete) directed information flow between variables, that is the causal influence. Such a definition can only be evaluated on the basis of some properties that we wish a measure of causal influence to have according to our "taste" or intuition. A list of such properties, that will be the axioms of partial information decomposition, is still under debate[Rau+14].

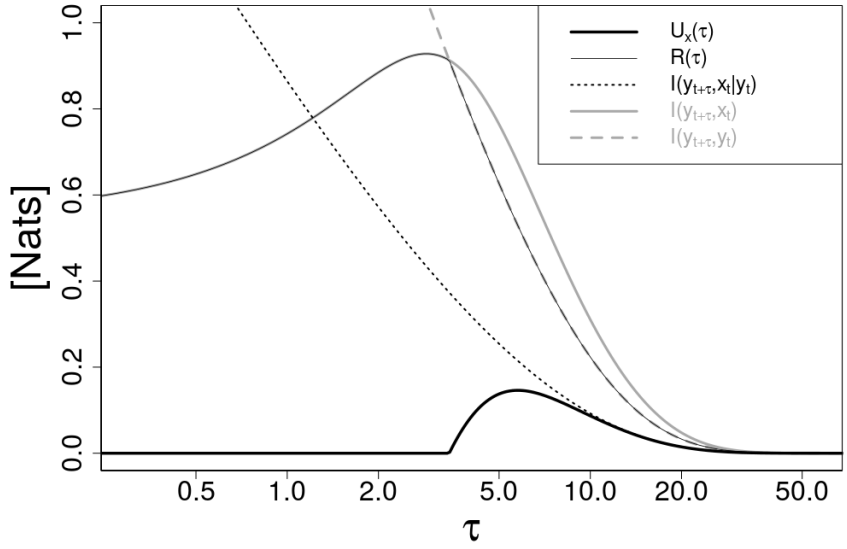


Fig. 4.3: Previously proposed PIDs where $R(\tau) = I_{min}$. The information measures are expressed in natural units $Nats = \frac{bits}{\ln 2}$. The τ axis is in logarithmic scale. The parameters are $\beta = 0.2$, $t_{rel} = 10$. Plot taken from [Auc+17].

In Gaussian systems[Bar15] the previously defined partial information decompositions (PIDs) are all equally just taking as redundancy the minimum value between $I(x(t), y(t+\tau))$ and $I(y(t), y(t+\tau))$, regardless of the information shared between the two sources $I(x(t), y(t))$, $I_{min} = \min[I(x(t), y(t+\tau)), I(y(t), y(t+\tau))]$. We plot the resulting PID and unique information in Fig.4.3, and we see that the unique information U_x is zero as long as $I(x(t), y(t+\tau))$ is smaller than $I(y(t), y(t+\tau))$. This form of activation is not in the dynamics, and is just mirroring the logic process of taking a minimum between two values. In addition, consider the bizarre effect in the point of activation of U_x, which corresponds to the time shift $\tau = \tau_e$ for which $I(x(t), y(t+\tau_e)) = I(y(t), y(t+\tau_e))$. At $\tau = \tau_e$ the two sources $x(t)$ and $y(t)$ are giving the same amount of information on $y(t+\tau)$, but they are not giving *the same information* as it is seen by the fact that they lead to different predictions in general. Then it is just wrong to consider such equal amount of information as redundant.

We define instead the **Redundancy** $R(\tau) \equiv R(x_t, y_t; y_{t+\tau})$ as a composition of the information shared between the two sources $x(t)$ and $y(t)$, $I_{xy} \equiv I(x(t), y(t))$, and the total information that they share with the target $y(t+\tau)$, $I_{tot} \equiv I_{tot}(\tau) \equiv I(y(t+\tau), (x(t), y(t)))$:

$$R(\tau) \equiv \frac{1}{2} \ln \left(\frac{e^{2(I_{xy}+I_{tot})}}{e^{2I_{xy}} + e^{2I_{tot}} - 1} \right).$$

that is Eq.(4.2) here, and Eq.(2) in the paper.

4.2.1 The idea behind the definition

The definition (4.2) was motivated by the analogy with the information propagation in a linear Markov chain. We detail here what is just stated in the paper [Auc+17] about this. Consider the *static* Gaussian linear network $A \to B \to C$, defined by the Bayesian structure:

$$\begin{cases} A = N_A \\ B = \gamma_B A + N_B \\ C = \gamma_C B + N_C \end{cases} \tag{4.22}$$

where the Ns are Gaussian Random variables with zero expectation and variance $\langle N^2 \rangle = 1$. This and the linear property make the expectations vanish: $\langle A \rangle = \langle B \rangle = \langle C \rangle = 0$. The variances are easily calculated as $\sigma_A^2 = 1$, $\sigma_B^2 = \gamma_B^2 + 1$, and $\sigma_C^2 = \gamma_C^2(\gamma_B^2 + 1) + 1$. Then correlations like $C_{AB} = \frac{\langle AB \rangle}{\sigma_A \sigma_B}$ are calculated as:

$$\begin{cases} C_{AB} = \frac{\gamma_B}{\sqrt{\gamma_B^2+1}} \\ C_{BC} = \frac{\gamma_C(\gamma_B^2+1)}{\sqrt{\gamma_B^2+1}\sqrt{\gamma_C^2(\gamma_B^2+1)+1}} \\ C_{AC} = \frac{\gamma_B \gamma_C}{\sqrt{\gamma_C^2(\gamma_B^2+1)+1}} \end{cases} \tag{4.23}$$

Using The Gaussian relation between mutual information and correlations $I = -\frac{1}{2}\ln(1 - C^2)$ we find the relation:

$$I_{AC} = I_{AB} + I_{BC} - \frac{1}{2}\ln\Big(e^{2I_{AB}} + e^{2I_{BC}} - 1\Big), \tag{4.24}$$

that is the form we choose for the definition of our Redundancy measure (4.2). Translated into (4.2) we defined the redundancy as the information that $x(t)$ *virtually* has on $y(t+\tau)$ when considering $x(t) \to y(t+\tau)$ as a linear channel with $y(t)$ in between. It is already clear that such definition is inappropriate for nonlinear systems, and it is even more problematic in feedback systems as we will discuss.

Let us try to motivate our choice also from another perspective. Let us state again clearly that the definition of Redundancy $R(\tau) \equiv \frac{1}{2}\ln\left(\frac{e^{2(I_{xy}+I_{tot})}}{e^{2I_{xy}}+e^{2I_{tot}}-1}\right)$, that is the main contribution of this chapter, is not derived from elementary principles. The logic behind this choice is found in the information processing properties of the BLRM. In particular, because of the absence of feedback, the information $I(y_t, x_{t+\tau})$ that y_t gives on the evolution of the signal $x_{t+\tau}$ can be expressed as a composition of the two consecutive information flows $I(y_t, x_t)$ and $I(x_t, x_{t+\tau})$:

$$I(y_t, x_{t+\tau}) = \frac{1}{2}\ln\left(\frac{e^{2(I(y_t,x_t)+I(x_t,x_{t+\tau}))}}{e^{2I(y_t,x_t)} + e^{2I(x_t,x_{t+\tau})} - 1}\right), \tag{4.25}$$

where we wrote the mutual information as $I(y_t, x_t)$ instead of $I(x_t, y_t)$ exactly to make it visible that the information $I(y_t, x_{t+\tau})$ flows through the chain $y_t \to x_t \to x_{t+\tau}$. We want the causal influence $C_{y\to x}(\tau)$ of the response on the signal to be zero. This is easily obtained defining the redundant information with the same functional form of (4.25). Translated into the causal influence $C_{x\to y}(\tau)$ it would mean to define the redundancy in the alternative form:

$$R^{alt}(\tau) \equiv \frac{1}{2}\ln\left(\frac{e^{2(I_{xy}+I(y_t,y_{t+\tau}))}}{e^{2I_{xy}}+e^{2I(y_t,y_{t+\tau})}-1}\right). \tag{4.26}$$

Now $I(x_t, x_{t+\tau})$ is equivalent to $I((x_t, y_t), x_{t+\tau})$ in signal-response models, and one obtains zero causal influence $C_{y\to x}(\tau)$ also with the functional form of (4.1), that is our definition. While $C_{y\to x}(\tau) = C^{alt}_{y\to x}(\tau) = 0$, the difference between the two definitions is observed in the causal influence of the signal on the response where $C_{x\to y}(\tau) \leq C^{alt}_{x\to y}(\tau)$.

Then one might ask: Why do we prefer to take the total information $I_{tot} \equiv I((x_t, y_t), y_{t+\tau})$ instead of $I(y_t, y_{t+\tau})$ in the definition of redundancy? The answer lies in a ***symmetry*** requirement: x_t and y_t are two sources for which a priori we don't want to give any preference in the prediction of $y_{t+\tau}$. The only artificial distinction we want to make is the one between sources and targets, that is to fix the direction of the time arrow considering the variables at time t as causes to those at time $t+\tau$. Then the symmetry between x_t and y_t requires us to define the redundancy measure as a composition between the mutual information $I_{xy} = I(x_t, y_t)$ and the total information that the two sources together give on the target $I_{tot} = I((x_t, y_t), y_{t+\tau})$. Importantly, the functional form (4.1)-(4.2) is also symmetric in I_{xy} and I_{tot} and can not exceed I_{xy}.

4.2.2 Causal influence properties in the BLRM

The **causal influence** resulting from the Redundancy definition 4.2,

$$C_{x\to y}(\tau) = I(x_t, y_{t+\tau}) - R(x_t, y_t; y_{t+\tau}), \tag{4.27}$$

is plotted in fig.4.4-4.5 for the BLRM. Here is the explicit expression as a function of the parameters t_{rel} and β:

$$C_{x\to y}(\tau) = \tfrac{1}{2}\ln\left(\frac{\sigma_y^2}{\sigma^2_{y_{t+\tau}|x_t}}\right) - \tfrac{1}{2}\ln\left(\frac{\sigma^2_{y_{t+\tau}|x_t,y_t}}{\sigma_y^2} + \frac{\sigma^2_{y|x}}{\sigma_y^2} - \frac{\sigma^2_{y_{t+\tau}|x_t,y_t}}{\sigma_y^2}\frac{\sigma^2_{y|x}}{\sigma_y^2}\right) =$$
$$= \tfrac{1}{2}\ln\left(\frac{(\beta t_{rel}-1)^2 - \frac{\beta t_{rel}}{\beta t_{rel}+1}\left(\beta t_{rel}(\beta t_{rel}+1)e^{-\frac{2\tau}{t_{rel}}} - 4\beta t_{rel}e^{-\tau(\beta+\frac{1}{t_{rel}})} + (\beta t_{rel}+1)e^{-2\beta\tau}\right)}{(\beta t_{rel}-1)^2 - \frac{\beta t_{rel}}{\beta t_{rel}+1}\left((\beta t_{rel}+1)^2 e^{-\frac{2\tau}{t_{rel}}} - 4(\beta t_{rel}+1)e^{-\tau(\beta+\frac{1}{t_{rel}})} + 4e^{-2\beta\tau}\right)}\right). \tag{4.28}$$

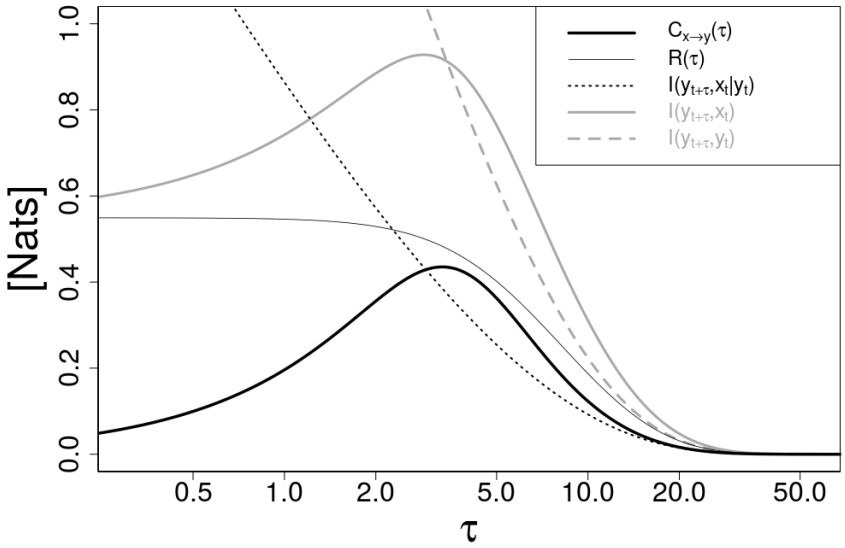


Fig. 4.4: Linear information decomposition $x \longrightarrow y$. In thick black is the unique information that $x(t)$ gives on $y(t+\tau)$, that is our measure of causal influence $C_{x\to y}(\tau)$. The parameters are $\beta = 0.2$, $t_{rel} = 10$. The plot is taken from [Auc+17].

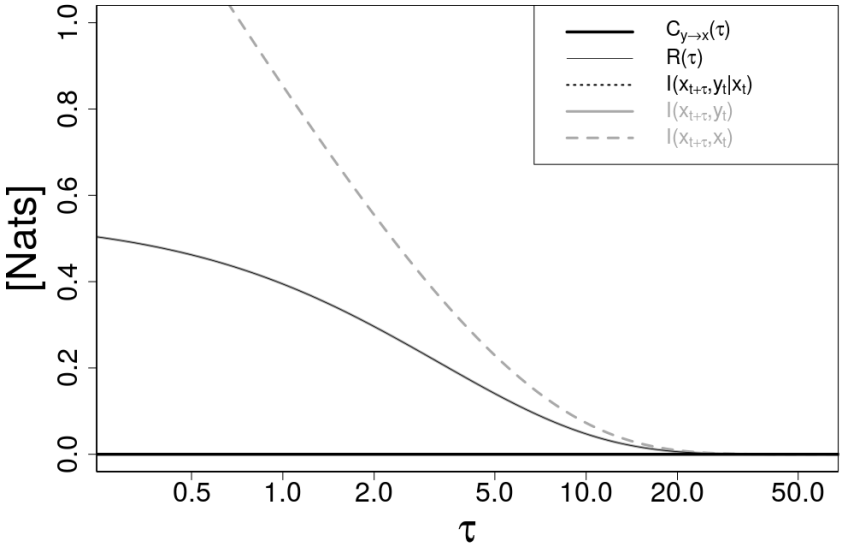


Fig. 4.5: Linear information decomposition $y \longrightarrow x$. The redundant information is equal to the mutual information meaning that there's no causal influence. The parameters are $\beta = 0.2$, $t_{rel} = 10$. The plot is taken from [Auc+17].

We force the definition of causal influence $C_{x\to y}(\tau)$ to only consider positive $\tau \geq 0$ as a postulate that effects are always seen after causes. Interestingly, one would get negative values for $C_{x\to y}(-\tau)$ (with $\tau > 0$) in the BLRM. This is unavoidable since the derivative of $C_{x\to y}(\tau)$ is continuous. In linear Langevin networks without feedback, and for $\tau \geq 0$, negative values of causal influence are never found. The main difference of our information decomposition to the previously defined ones is that our redundancy R is explicitly dependent on the information shared between the two sources giving the redundant information on the target and is always less or equal to that. Note that $C_{x\to y}(\tau)$ is measured in $Nats$ like all other information measures.

Let us discuss the curves in Fig.4.4-4.5. The absence of noise in the response implies that the knowledge of the continuous history of x for a sufficiently long time allows the determination of y with any desired precision. We can say that the value of y is caused by the trajectory of x up to that time. As a result of this, the transfer entropy is diverging for small time shifts $\tau \to 0$. Nevertheless we defined[Auc+17] as causes the single observable facts ($x(t)$ and $y(t)$ in the BLRM), and as the effect a successive observable fact ($y(t+\tau)$), and causal influences quantify the relative strength of these causes in giving the effect.

The causal influence $C_{x\to y}(\tau)$ that the signal has on the response over time starts from 0 at $\tau = 0$, and then increases with τ (linearly for small τ) reflecting the fact that the effect of causality is seen gradually over time after the cause $x(t)$. For very long time intervals τ after the cause we cannot see anymore the effect of the distant past and the causal influence goes to 0. The time shift at which the causal influence peaks τ_{res} is the response time of the system in the probabilistic sense and is slightly different from the maximum correlation time τ_{opt}. We find that $\tau_{res} > \tau_{opt}$.

As we already discussed, we get zero causal influence of the response y on the signal x (fig.4.5), and that is because the information $I(y(t), x(t+\tau))$ that the response has on the evolution of the signal is gained necessarily via the two steps $y(t) \to x(t)$ and $x(t) \to x(t+\tau)$ due to the asymmetry of the interaction (no feedback) and therefore is equal to the redundancy $R(x(t), y(t); x(t+\tau))$. Let us just mention that the self-causal influence is zero, $C_{x\to x} = 0$ and $C_{y\to y} = 0$, meaning that causation is exerted only between different observables, while the autocorrelation of one variable just means memory of the previous states.

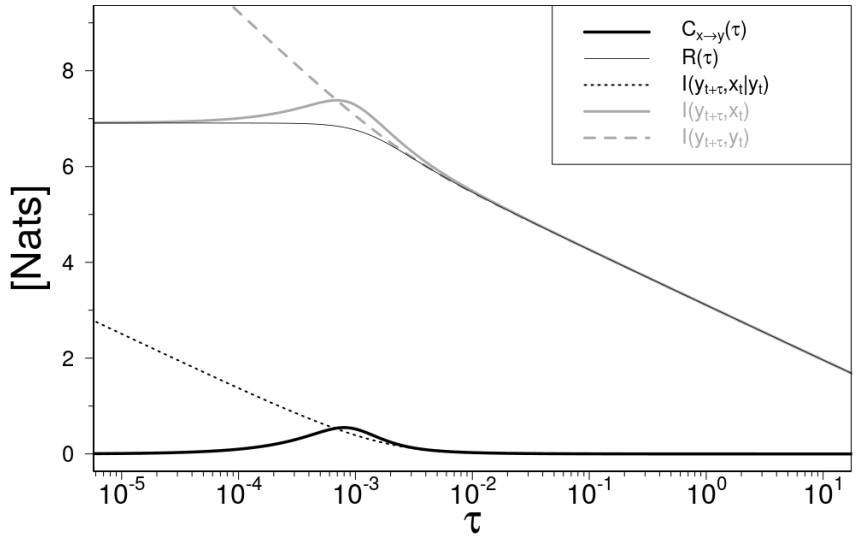


Fig. 4.6: Linear information decomposition $x \longrightarrow y$. High information scenario: $\beta = 1000$, $t_{rel} = 1000$. The plot is taken from [Auc+17].

The concepts of redundancy and synergy were originally defined (outside of the PID framework) as a single quantity, the "net redundancy/synergy" coinformation measure[Sch+03]:

$$\begin{gathered}\mathbf{CoI}(y(t+\tau), x(t), y(t)) \equiv \\ I((x_t, y_t), y_{t+\tau}) - I(y(t+\tau), x(t)) - I(y(t+\tau), y(t)) = \\ I(y(t+\tau), x(t)|y(t)) - I(y(t+\tau), x(t)).\end{gathered} \tag{4.29}$$

Positive and negative values of CoI indicate respectively synergy and redundancy. The BLRM is synergistic for small time shifts τ and redundant for larger τ when the mutual information exceeds the transfer entropy. **CoI** is symmetric in its three arguments $(x(t), y(t), y(t+\tau))$ and the relation with the PID measures is simply $\mathbf{CoI}(y(t+\tau); x(t), y(t)) = S - R$. The coinformation measure was not useful in our definition of causal influence, and we reported it for the sake of completeness only.

4.2.3 Parameter study and asymptotic behavior

To understand the behavior of the causal influence in the BLRM as a function of the parameters we study the limits of high and low information (fig.4.6-4.7). The following list of properties is just quoted from our paper [Auc+17]. When $\beta t_{rel} >> 1$ the mutual information is high and increases with $\ln(\beta t_{rel})$. The peak of the causal influence also increases but only up to a limit of around $\lim_{\beta t_{rel}\to\infty} \max_\tau C_{x\to y} \approx 0.55\ Nats$. The position of the peak depends on β: with higher β the response is faster and the effect of causality is seen earlier. When β is fixed, increasing t_{rel} gives always an increase in the mutual information because the slow down of the dynamics

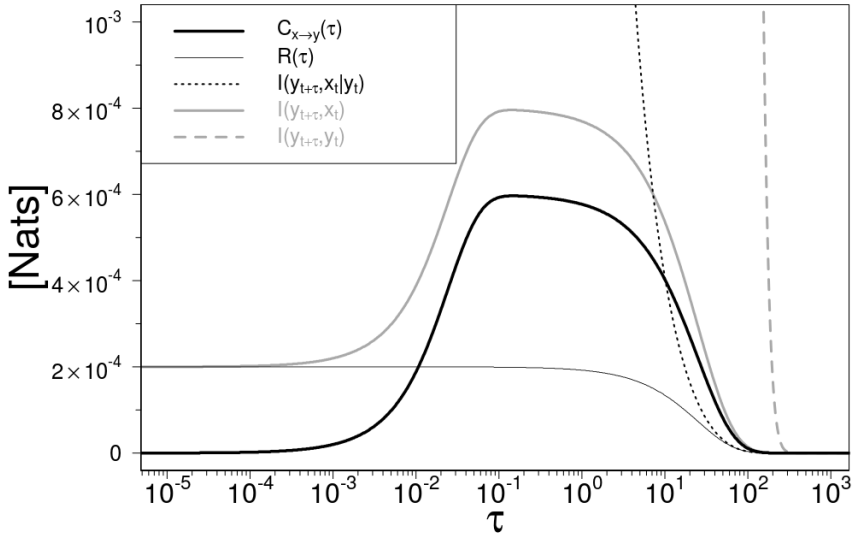


Fig. 4.7: Linear information decomposition $x \longrightarrow y$. Low information scenario: $\beta = 0.02$, $t_{rel} = 0.02$. The plot is taken from [Auc+17].

of the signal lets the response follow the microscopic structure of the signal with more precision, but at the same time the response is moved slower (in units of his standard deviation) by the signal, these two effects asymptotically compensating and the peak of the causal influence staying around $\approx 0.55\ Nats$. This limit we call the causation capacity of the BLRM. In the case of low information $\beta t_{rel} << 1$ the peak of the causal influence is close to 75% of the peak of the mutual information because $\frac{I^{opt}}{I_{xy}} \to 4$ for $\beta t_{rel} \to 0$. The signal has fast-decaying autocorrelation, the response is slowly integrating (keeping the memory of) it and therefore most of the small amount of time-lagged mutual information on the response is causal influence.

4.2.4 Comparison with vector autoregressive models

Let us now consider a more traditional approach in data analysis. We write the vector autoregressive model (VAR) for the evolution of the response as $y(t+\tau) = \gamma_{yy}(\tau)y(t) + \gamma_{xy}(\tau)x(t) + \xi(\tau)$, where the γs are the linear coefficients of the expansion and $\xi(\tau)$ is the error term. One could consider the coefficient $\gamma_{xy}(\tau) = \frac{\alpha t_{rel}}{\beta t_{rel}-1}(e^{-\frac{\tau}{t_{rel}}} - e^{-\beta\tau})$ as a measure of the influence of the signal on the response, but then the error term $\xi(\tau)$ would have *no explicit* role. The intuition of causal influence is based on the concept of information flow and the fact that the uncertainty on the prediction $\langle \xi^2(\tau) \rangle = \sigma^2_{y(t+\tau)|x(t),y(t)}$ increases with τ should consequently decrease the causal influence. The discrepancy of the coefficient $\gamma_{xy}(\tau)$ with the information measures is clearly seen in fig.4.8.

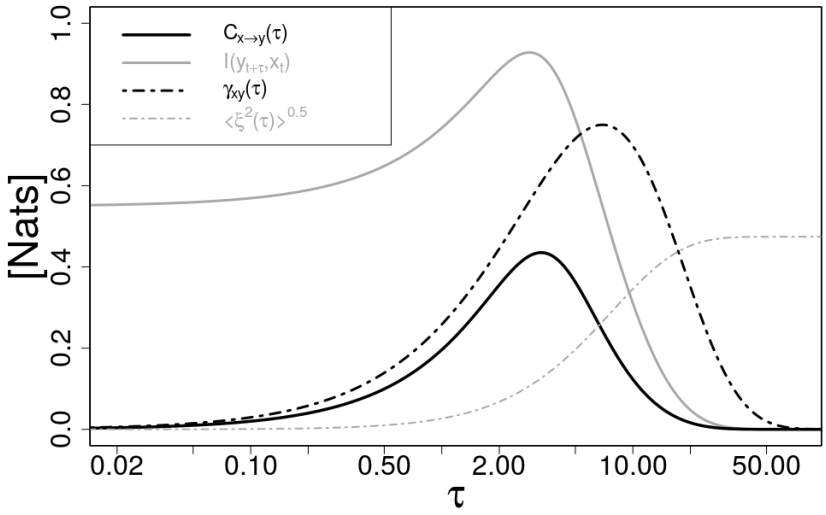


Fig. 4.8: The coefficient γ_{xy} of the vector autoregressive model compared with the information measures. γ_{xy} and $\sqrt{\langle \xi^2(\tau) \rangle}$ are adimensional. The parameters are $\beta = 0.2$, $t_{rel} = 10, \alpha = 0.3, D = 0.03$. The plot is taken from [Auc+17].

One dissatisfying feature of our definition, however, is that the redundancy measure does not satisfy the local positivity axiom of the PID, i.e. the synergistic information $S = T_{x \to y} - C_{x \to y}$ is negative when the causal influence is greater than the transfer entropy, and this is always the case for long delays τ. This means that part of the "same" information that $x(t)$ and $y(t)$ give on $y(t+\tau)$ is considered as causal influence and not redundancy. We add here to the paper that the axioms proposed for partial information decompositions by different authors were demonstrated[Rau+14] to be non compatible. In particular the local positivity is in contrast with the basic axioms of Williams-Beer[WB10].

4.3 Multidimensional case: networks without feedbacks

We can extend the causal influence measure for interactions within multidimensional linear Langevin networks without feedbacks. Let us define the network of direct influences as the one that has directed links for all the combination of variables (nodes in the network) $(i \to j)$ for which variable i appears in the equation for the dynamics of variable j. The network of the causal influence is not coincident with the network of direct influences because we also have to consider as causal all the indirect influences. Let us define the parents P_x of a node x as the set of all nodes in the network of direct influences that are able to reach x with directed paths. We expect all the parents P_x to have causal influence on x, in general with different intensities and time scales. Similarly, we define the common parents P_{xy} of two

nodes x and y as the set of all nodes in the network of direct influences that are able to reach both nodes x and y with directed paths.

Then we generalize the definition of **causal influence** to the **multidimensional case** adding the condition of the knowledge of the state of the common parents $P_{xy}(t)$ at time t to all the probability measures:

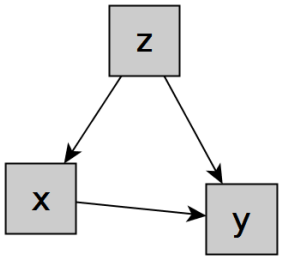
$$C_{x\to y}(\tau) = I(x(t), y(t+\tau)|P_{xy}(t)) - R(x(t), y(t); y(t+\tau)|P_{xy}(t)), \tag{4.30}$$

where $R(x(t), y(t); y(t+\tau)|P_{xy}(t))$ is defined as in eq.(4.2) but with all the information measures conditioned to the knowledge of the state of the common parents $P_{xy}(t)$ at time t.

4.3.1 Feed-forward loop

For simplicity we consider as an example the network of three nodes without feedbacks, that is the so called feed-forward loop:

$$\begin{cases} \frac{dz}{dt} = -\frac{z}{t_{rel}} + \sqrt{D_z}\,\Gamma_z(t) \\ \frac{dx}{dt} = \alpha_x z - \beta_x x + \sqrt{D_x}\,\Gamma_x(t) \\ \frac{dy}{dt} = \alpha_y z - \beta_y y + \gamma x + \sqrt{D_y}\,\Gamma_y(t) \end{cases} \tag{4.31}$$



When the $x \to y$ interaction parameter is zero, $\gamma = 0$, the variable x is not a parent of y and therefore it should have no causal influence on it. Still, x and y can be highly correlated due to the common parent z. Applying the above definition we analytically calculated the causal influence of x on y and it results to be zero, $C_{x\to y}(\tau) = I(x(t), y(t+\tau)|z(t)) - R(x(t), y(t); y(t+\tau)|z(t)) = 0$. The analytical calculation is done in the Appendix A of [Auc+17] and is commented here.

We study the particular case of the system of eq.4.31 without influence of the x on the y, i.e. with $\gamma = 0$. We calculate here all the information measures needed to show that the causal influence $C_{x\to y}(\tau)$ of system (4.31) with $\gamma = 0$ is zero. We start from those quantities which are found also in the BLRM. The conditional expectation

values and the standard deviations for the couples zx and zy are symmetric so we write them just once:

$$\begin{aligned}
&\langle z(t-\tau+t')z(t-\tau+t'+t'')|z(t)\rangle = \\
&= \int_{-\infty}^{+\infty} P(z(t-\tau+t'+t'')=\xi|z(t))\xi \,\langle z(t-\tau+t')|z(t-\tau+t'+t'')=\xi\rangle \, d\xi = \\
&= e^{-\frac{t''}{t_{rel}}}\left(z^2(t)e^{-\frac{2(\tau-t'-t'')}{t_{rel}}} + \sigma_z^2(1-e^{-\frac{2(\tau-t'-t'')}{t_{rel}}})\right),
\end{aligned} \tag{4.32}$$

where we used the Markov property for z, and the Chapman-Kolmogorov equation[Gar09]. x has no direct influence on z, then it holds $p(x(t)|z(t), z(t+\tau)) = p(x(t)|z(t))$, and we obtain:

$$\begin{aligned}
&\langle z(t-\tau+t')x(t-\tau+t')|z(t)\rangle = \\
&= \int_{-\infty}^{+\infty} P(z(t-\tau+t')=\xi|z(t))\xi \,\langle x(t-\tau+t')|z(t-\tau+t')=\xi\rangle \, d\xi = \\
&= \tfrac{\alpha_x t_{rel}}{\beta_x t_{rel}+1}(z^2(t)e^{-\frac{2(\tau-t')}{t_{rel}}} + \sigma_z^2(1-e^{-\frac{2(\tau-t')}{t_{rel}}})),
\end{aligned} \tag{4.33}$$

$$\begin{aligned}
&\langle z(t-\tau+t')x(t)|z(t)\rangle = \\
&= \langle z(t-\tau+t')x(t-\tau+t')|z(t)\rangle \, e^{-\beta_x(\tau-t')} + \\
&+ \quad \alpha_x \int_0^{\tau-t'} \langle z(t-\tau+t')z(t-\tau+t'+t'')|z(t)\rangle \, e^{-\beta_x(\tau-t'-t'')}dt'' = \\
&= \sigma_z^2 \tfrac{2\alpha_x t_{rel}}{\beta_x^2 t_{rel}^2-1}(e^{-\frac{\tau-t'}{t_{rel}}} - e^{-\beta_x(\tau-t')}) + z^2(t)\tfrac{\alpha_x t_{rel}}{\beta_x t_{rel}+1}e^{-\frac{\tau-t'}{t_{rel}}},
\end{aligned} \tag{4.34}$$

The variances and conditional variances of y have the additional term $\frac{D_y}{2\beta_y}$ compared to the BLRM that is due to the noise source Γ_y:

$$\sigma_y^2 = \sigma_z^2 \frac{\alpha_y^2 t_{rel}}{\beta_y(1+\beta_y t_{rel})} + \frac{D_y}{2\beta_y} \tag{4.35}$$

$$\sigma^2_{y(t+\tau)|z(t)} = \sigma_y^2 - (\frac{\sigma_z \alpha_y t_{rel}}{\beta_y t_{rel}-1})^2(e^{-\frac{\tau}{t_{rel}}} - \frac{2e^{-\beta_y\tau}}{\beta_y t_{rel}+1})^2, \tag{4.36}$$

$$\begin{aligned}
&\sigma^2_{y(t+\tau)|x(t),y(t),z(t)} = \sigma^2_{y(t+\tau)|y(t),z(t)} = \tfrac{\sigma_z^2\alpha_y^2}{\beta_y(\beta_y+1/t_{rel})(\beta_y-1/t_{rel})^2} * \\
&*[(1-\beta_y t_{rel})^2 - e^{-2\beta_y\tau}(1+\beta_y t_{rel}) + e^{-(\beta_y+\frac{1}{t_{rel}})\tau}4\beta_y t_{rel} - e^{-\frac{2\tau}{t_{rel}}}\beta_y t_{rel}(1+\beta_y t_{rel})] + \\
&+\tfrac{D_y}{2\beta_y}(1-e^{-2\beta_y\tau}),
\end{aligned} \tag{4.37}$$

$$\langle y(t-\tau)x(t)|z(t)\rangle = \langle y(t)x(t)|z(t)\rangle \, e^{\beta_y\tau} - \alpha_y \int_0^\tau \langle z(t-\tau+t')x(t)|z(t)\rangle \, e^{\beta_y t'}dt'. \tag{4.38}$$

$\langle y(t-\tau)x(t)|z(t)\rangle \to 0$ for $\tau \to \infty$ because the knowledge of $x(t)$ gives asymptotically no information on the distant past of y (even with the condition $z(t)$), then using (4.34) we obtain:

$$\langle y(t)x(t)|z(t)\rangle = \frac{\alpha_x \alpha_y t_{rel}^2}{(\beta_x t_{rel}+1)(\beta_y t_{rel}+1)}(z^2(t) + \frac{2\sigma_z^2}{t_{rel}(\beta_x+\beta_y)}). \tag{4.39}$$

Since $\langle z(t+t')x(t)|z(t)\rangle = \langle z(t+t')|z(t)\rangle \langle x(t)|z(t)\rangle$, whose quantities we discussed already in the BLRM, we can now calculate $\langle y(t+\tau)x(t)|z(t)\rangle = \langle y(t)x(t)|z(t)\rangle e^{-\beta_y \tau} + \alpha_y \int_0^\tau \langle z(t+t')x(t)|z(t)\rangle e^{-\beta_y(\tau-t')}dt'$ and then the correlation:

$$\langle y(t+\tau)x(t)|z(t)\rangle - \langle y(t+\tau)|z(t)\rangle \langle x(t)|z(t)\rangle = \sigma_z^2 \frac{2\alpha_x\alpha_y t_{rel} e^{-\beta_y\tau}}{(\beta_x t_{rel}+1)(\beta_y t_{rel}+1)(\beta_x+\beta_y)}, \tag{4.40}$$

which is independent of the condition $z(t)$, as it is typically the case for linear systems. The information measures are easily calculated in the Gaussian case:

$$I_{tot} = \ln\left(\frac{\sigma^2_{y(t+\tau)|z(t)}}{\sigma^2_{y(t+\tau)|x(t),y(t),z(t)}}\right), \tag{4.41}$$

$$C(x(t), y(t+\tau)|z(t)) = \frac{\langle y(t+\tau)x(t)|z(t)\rangle - \langle y(t+\tau)|z(t)\rangle \langle x(t)|z(t)\rangle}{\sigma_{y(t+\tau)|z(t)}\sigma_{x(t)|z(t)}}, \tag{4.42}$$

$$I(x(t), y(t+\tau)|z(t)) = -\frac{1}{2}\ln\Big(1 - C^2(x(t), y(t+\tau)|z(t))\Big). \tag{4.43}$$

Using the definition of redundancy (4.2), $R(\tau) = \frac{1}{2}\ln\Big(\frac{e^{2(I_{xy}+I_{tot})}}{e^{2I_{xy}}+e^{2I_{tot}}-1}\Big)$, with $I_{xy} = -\frac{1}{2}\ln(1 - C^2(x(t), y(t)|z(t)))$ we obtain the expected result:

$$C_{x\to y}(\tau) = I(x(t), y(t+\tau)|z(t)) - R(x(t), y(t); y(t+\tau)|z(t)) = 0. \tag{4.44}$$

When $\gamma \neq 0$ our causal influence measure $C_{x\to y}(\tau)$ would detect the presence of the $x \to y$ influence (numerical results in fig.4.9). The shape of $C_{x\to y}(\tau)$ is qualitatively the same as in the BLRM. We verified numerically that the causal influence is correctly zero for the cases $C_{y\to x} = C_{x\to z} = C_{y\to z} = 0$. The transfer entropy $T_{x\to y}(\tau) = I(y(t+\tau), x(t)|y(t), z(t))$ goes to 0 for $\tau \to 0$ because the white noise term $\sqrt{D_y}\,\Gamma_y$ dominates the dynamics for short intervals.

Importantly, the fact of having a very small (negligible) direct interaction $z \to x$, i.e. $\alpha_x \ll \alpha_y$, implies that the probability distributions in the calculation of the causal influence $C_{x\to y}$ have to be conditioned on the common parent $z(t)$. On the contrary, without a direct interaction $z \to x$, i.e. $\alpha_x = 0$, z is not a common parent and therefore there's no conditioning on $z(t)$. However, as it should be, the

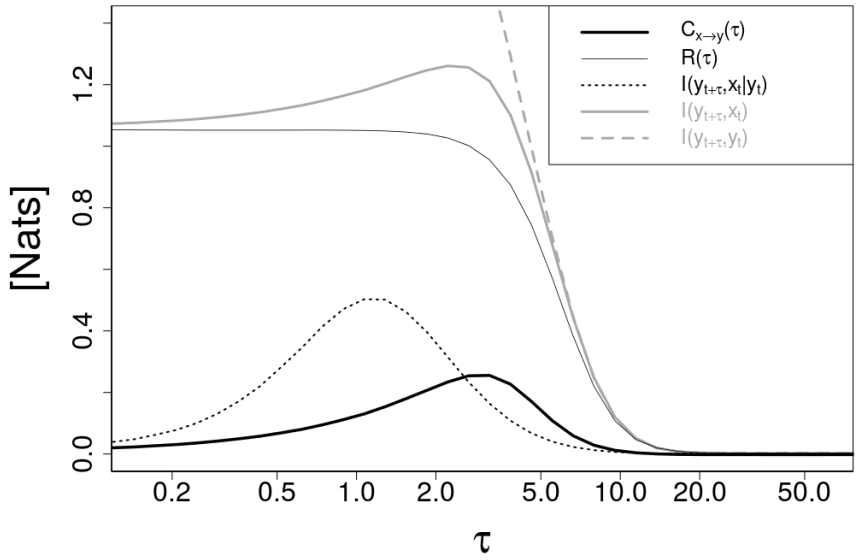


Fig. 4.9: Feed-forward loop, the 3-dimensional general case. Causal influence $x \longrightarrow y$ (numerical simulation). The parameters are $t_{rel} = 10$, $\gamma = \alpha_x = \alpha_y = 1$, $\beta_x = \beta_y = 0.2$, $D_z = 10$, $D_x = D_y = 0.1$. It is not explicitly written in the legend, but note that all the information measures are here conditioned on the common parent state $z(t)$ (see Eq.4.30). The plot is taken from [Auc+17].

conditioning makes no difference in the limit $\alpha_x \to 0$: we numerically verified that $\lim_{\alpha_x \to 0} C_{x\to y} = C_{x\to y}(\alpha_x = 0)$.

We note that even without a direct interaction $z \to y$, that is $\alpha_y = 0$, the causal influence $C_{z\to y}$ can be positive due to the indirect influence $z \to x \to y$. The bigger is the number of indirect passages between the considered nodes, the longer is the time period τ after which the peak of the causal influence is seen.

Here the conditioning on the common parents $P_{xy}(t)$ can be seen as a negative feature since it introduces again in the multidimensional case the synergistic effects, which are properties of the transfer entropy and of any conditioning[Jam+16]. However there are no other possibilities since the construction of a PID lattice as the one proposed by Williams and Beer[WB10] would require a generalization of our definition of redundancy $R(\tau)$ for more than two sources. For this we would need to have an expression for the mutual information I_{xy} between more than two variables, which is not defined in information theory[CT12].

Let us study the (conditional) causal influence (4.30) $C_{x\to y}(\tau)$ for different values of the coupling with the common parent z. For simplicity we choose $\alpha_x = \alpha_y = \alpha$, and simulate with α of many different orders of magnitude. As it is seen in Fig.4.10, the causal influence (4.30) does not change much, meaning that the effect of conditioning is to isolate the causal interaction $x \to y$, that would not be observed for $\alpha >> 1$ (and $\beta_x = \beta_y > 0$ and $D_x = D_y > 0$) without the knowledge of z.

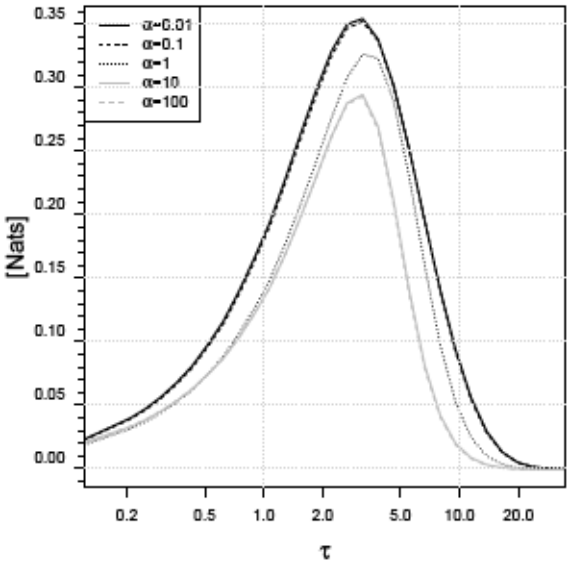


Fig. 4.10: Conditional causal influence $C_{x\to y}(\tau)$ for different values of the common parent interaction parameter $\alpha_x = \alpha_y = \alpha$. The other parameters are $\beta_x = \beta_y = 0.2$, $D_x = D_y = 0.01$, $D_z = 0.1$, $t_{rel} = 10$, and $\gamma = 1$.

Larger α makes the role of z being more important, and that is probably why the causal influence decreases. This suggest that a conditioning on the whole history of z between t and $t+\tau$ could be the most appropriate definition of conditional causal influence in the 3D case. That would be, anyway, computationally non feasible.

4.3.2 Competing influence

Let us now use the causal influence to measure how two competing inputs x and z determine the dynamics of a single target y. This is still a signal-response model, and can be written as:

$$\begin{cases} \frac{dx}{dt} = -\frac{x}{t_{rel}} + \sqrt{D_x}\,\Gamma_x(t) \\ \frac{dz}{dt} = -\frac{z}{t_{rel}} + \sqrt{D_z}\,\Gamma_z(t) \\ \frac{dy}{dt} = \alpha_x x + \alpha_z z - \beta y \end{cases} \tag{4.45}$$

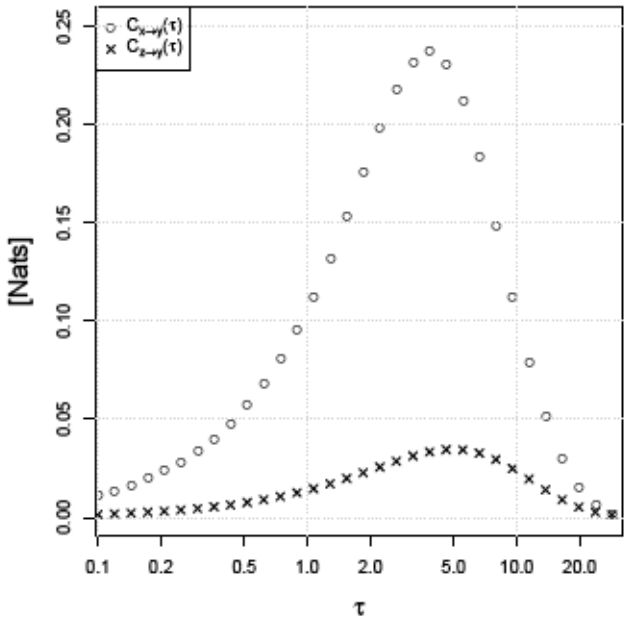
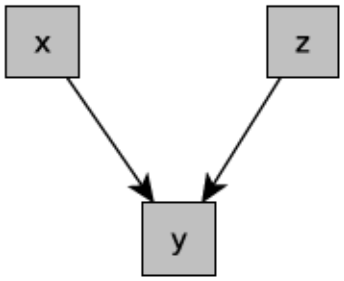


Fig. 4.11: Causal influences on variable y given by the two competing variables x and z. These are respectively $C_{x\to y}(\tau)$ and $C_{z\to y}(\tau)$. The parameters are $\beta = 0.2$, $D_x = D_z = 1$, $D_y = 0$, $t_{rel} = 10$, $\alpha_x = 1$, and $\alpha_z = \frac{\alpha_x}{2} = 0.5$.



For simplicity, we did not consider an intrinsic noise source for y. Now we consider the case in which the (asymmetric) interaction parameter $\alpha_z = \frac{\alpha_x}{2} = 0.5$, so that variable x has a larger impact on the dynamics of y. This is reflected on the causal influence intensity as it is seen in Fig.4.11, where the peak of $C_{x\to y}(\tau)$ is higher than the peak of $C_{z\to y}(\tau)$.

4.4 Feedback systems and difficulties in the generalization

Finally we discuss the strong limitation of our causal influence definition: there is currently no way of extending it to any system with general feedback structure.

First of all, it is not clear whether the concept of causal influence still makes sense in the presence of feedback. When the variable x is influencing the dynamics of the variable y and vice versa forming a feedback loop, we can't define anymore a signal and a response. The $x(t)$ at time t is influencing the evolution of the response $y(t+\tau)$ at time $t+\tau$ in many ways: directly and also indirectly through the loop $x(t) \to y(t+t') \to x(t+t'') \to y(t+\tau)$ with $\tau > t'' > t' > 0$, but also through the loops $x(t) \to y(t+t') \to x(t+t'') \to y(t+t''') \to x(t+t'''')... \to y(t+\tau)$ and so on. These "successive" influences are in opposing directions for linear negative feedback loops and this implies the information measures to oscillate over time and our measure of causal influence to oscillate as well and to assume also negative values. Since the mutual information $I(x(t), y(t+\tau))$ is periodically assuming 0, we would have oscillating causal influence with any definition of R and we may conclude that the point-to-point communication scheme $(t, t+\tau)$ is not appropriate for a definition of causal influence in the presence of feedbacks.

Then another option could be to take the original definition of transfer entropy (it considers the whole past history of the studied processes) for discrete time sequences[Sch00] and generalize it for continuous signals: $\mathbf{TE}^{seq}_{x\to y}(\tau) = I(y(t+\tau), (x(t-t'))_{t'\geq 0} | (y(t-t'))_{t'\geq 0})$, where $(x(t-t'))_{t'\geq 0}$ and $(y(t-t'))_{t'\geq 0}$ are the semi-infinite whole past history of the signal and response, respectively. For any set of stochastic differential equations without delays (also in the presence of feedbacks), if we already have the knowledge of the present values of the signal $x(t)$ and of the response $y(t)$, then the past of the signal $(x(t-t'))_{t'\geq 0}$ gives no additional information on the future of the response $y(t+\tau)$ and $\mathbf{TE}^{seq}_{x\to y}(\tau) = I(y(t+\tau), x(t) | (y(t-t'))_{t'\geq 0}) = I(y(t+\tau), x(t) | y(t), P(x(t) | (y(t-t'))_{t'\geq 0}))$. This function is difficult to estimate even in linear response models. However, in the BLRM, since the knowledge of $(y(t-t'))_{t'\geq 0}$ gives infinitely big amount of information on $x(t)$, the generalized transfer entropy would be zero, $\mathbf{TE}^{seq}_{x\to y}(\tau) = 0$, and this would be the case also for any other bi-dimensional model (also in the presence of feedback) with no intrinsic noise in the dynamics of the response. The BLRM is our prototype in which frame to quantify the evident influence that the signal has on the response and we would not be satisfied to identify the generalized transfer entropy $\mathbf{TE}^{seq}_{x\to y}(\tau)$, which is here always zero, as a measure of causal influence.

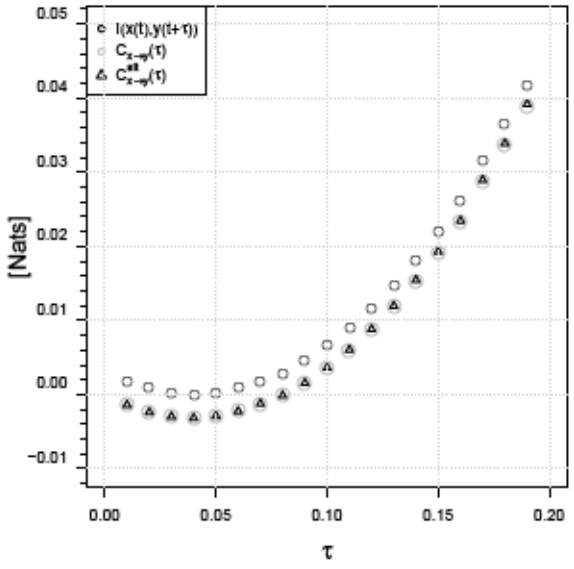


Fig. 4.12: This is to show how the causal influence would perform in a imbalanced feedback model like (4.46). Both $C_{x\to y}(\tau)$ and $C^{alt}_{x\to y}(\tau)$ are negative for $\tau \sim 0.05$. The parameters are $\beta = 0.2$ and $\alpha = 0.1$.

An additional problem in defining a measure of causal influence could be the partial information decomposition itself. What is the meaning of decomposing the mutual information? The mutual overlap between the two probability distributions $P(y(t+\tau)|y(t))$ and $P(y(t+\tau)|x(t))$ could be defined as

$$\left\langle \int_{-\infty}^{\infty} P(y(t+\tau)|y(t))P(y(t+\tau)|x(t))dy(t+\tau) \right\rangle_{x(t),y(t)}$$

but then it's not easy to say whether such quantity can be used to define a measure of overlap in the space of Shannon entropies, i.e. the common information (redundancy) of $I(y(t+\tau), y(t))$ and $I(y(t+\tau), x(t))$.

It is also important to note that the non-negativity property is violated in feedback systems for both definitions (4.1)-(4.2) and (4.26). Let us consider for example the following (unbalanced) stochastic negative feedback loop:

$$\begin{cases} \frac{dx}{dt} = -\beta x - 2\pi y + \Gamma_x \\ \frac{dy}{dt} = -\beta y + \alpha 2\pi x + \Gamma_y \end{cases} \tag{4.46}$$

where Γ_x and Γ_y are white noise sources. Choosing the parameters $\beta = 0.2$ and $\alpha = 0.1$, both the causal influence definitions are negative for short intervals $\tau \sim 0.05$ as it is seen in Fig.4.12. The non-negativity is one dissatisfying feature of our partial

information decomposition. However, it was shown in[Rau+14] that the axioms of Willams-Beer (identity, symmetry, monotonicity) are not compatible with the local positivity axiom.

Let us say something about the previously proposed Redundancy measures[Ber+14; Gri+14]. While they have quite distinct definitions, they share two fundamental points. First, they consider only the marginal distribution of each individual source (x_t and y_t) with the target ($y_{t+\tau}$), regardless of the actual joint probability $p(x_t, y_t, y_{t+\tau})$, and more importantly, regardless of the information that the two sources share, $I(x_t, y_t)$. Second, they are defined in algorithmic way as a minimization of some quantity.

The main result of Ref.[Bar15] is that, in Gaussian systems, there is only one such partial information decomposition, and is the one that takes as redundancy the minimum between $I(x_t, y_{t+\tau})$ and $I(y_t, y_{t+\tau})$. This is what motivated us to search for a definition that takes the mutual information between the sources as a starting point and an upper bound to the redundancy.

We speculate that our definition could be a linear non-feedback approximation of a more general definition to be found.

Additional references to the paper [Auc+17] we did not mention regard the use of dynamic stochastic models and information theory in biology. They are listed here: [Kho06; Kir+16; BS05; Sch+03; DTW12; DTW14].

5 Information thermodynamics on time series

> *The irreversibility of time is the mechanism that brings order out of chaos.*
>
> **— Ilya Prigogine**

This chapter contains our original contribution to the field of Information Thermodynamics, that is the main result of this PhD thesis. The material presented is an important modification to our ArXiv manuscript [Auc+18], and it has recently been reviewed and published in [Auc+19b] (the first official version of this thesis was submitted before, in November 2018). Such publication contains therefore very similar results to the ones presented here. Anyway additional comments and more speculations on future research are presented here and not there.

5.1 Introduction

The irreversibility of a process is the possibility to infer the existence of a time's arrow looking at an ensemble of realizations of its dynamics[Jar11; Par+09; FC08]. This concept appears in the nonequilibrium thermodynamics quantification of dissipated work or entropy production[Jar97; Cro99; ES02], and it relates the probability of paths with their time-reversal conjugates[Kaw+07].

Fluctuation theorems have been developed to describe the statistical properties of the entropy production and its relation to information-theoretic quantities in both Hamiltonian and Langevin dynamics[Jar00; Che+06; IS13], and we largely discussed the main results in the field in Chapter 3. Still the reading of previous chapters is not essential to understand this Chapter, since a short selection of previous results relevant to this study are given hereafter. Particular attention was given to measurement-feedback controlled models[SU12; SU10] inspired by the Maxwell's demon[Szi64], a gedanken-experiment in which mechanical work is extracted from thermodynamic systems using information. An ongoing experimental effort is put in the design and optimization of such information engines[Mar+16; Cil17; Toy+10; Kos+14].

Theoretical studies clarified the role of fluctuations in feedback processes described by bipartite (or multipartite) stochastic dynamics, where fluctuation theorems set lower bounds on the entropy production of subsystems in terms of the Horowitz-Esposito information flow[HE14; RH16; Hor15], or in terms of the transfer entropy[Ito16; Spi+16] in the interaction with other subsystems.

Time series are obtained measuring a continuous underlying dynamics at a finite frequency $\frac{1}{\tau}$, and this is the case of most real data. A measure of irreversibility for time series was defined in [RP12] as the Kullback-Leibler divergence[CT12] between probability densities of time series realizations and of their time-reversal conjugates. Time series are non bipartite in general, and this prevents an unambiguous identification of the physical entropy production or heat exchanged with thermal baths[Sek98]. Then the time series irreversibility does not generally converge to the physical entropy production in the limit of high sampling frequency $\tau \to 0$, except for special cases like Langevin systems with constant diffusion coefficients, as we will discuss here.

The irreversibility measure depends on the statistical properties of the whole time series for non-Markovian process. We will consider a measure of irreversibility that considers only the statistics of single transitions, and that recovers the irreversibility of the whole trajectories only for Markovian systems. We call it *mapping irreversibility* and use the symbol Φ_τ.

We study fluctuations of the mapping irreversibility introducing its stochastic counterpart, φ_τ. In the bivariate case of two interacting variables x and y, we define the conditional stochastic mapping irreversibility $\varphi_\tau^{y|x}$ as the difference between that of the joint process and that of a single subsystem, $\varphi_\tau^{y|x} = \varphi_\tau^{xy} - \varphi_\tau^{x}$.

We define signal-response models as continuous-time stationary stochastic processes characterized by the absence of feedback. In the bidimensional case, a signal-response model consist of a fluctuating signal x and a dynamic response y. In chapter 4 (and in our paper [Auc+17]) we studied the information processing properties of linear multidimensional signal-response models. In that framework we defined a measure of causal influence to quantify how the macroscopic effects of asymmetric interactions are seen over time.

The backward transfer entropy $T_{y\to x}(-\tau)$ is the standard transfer entropy[CT12] calculated in the ensemble of time-reversed trajectories. It was already shown to have a role in stochastic thermodynamics, in gambling theory, and in anti-causal linear regression models[Ito16].

We derive an integral fluctuation theorem for time series of signal-response models that involves the backward transfer entropy. From this follows the II law of thermodynamics for signal-response models, i.e. that the backward transfer entropy $T_{y\to x}(-\tau)$ of the response y on the past of the signal x is a lower bound to the conditional mapping irreversibility $\Phi_\tau^{y|x} = \langle \varphi_\tau^{y|x} \rangle$:

$$\Phi_\tau^{y|x} \geq T_{y\to x}(-\tau). \tag{5.1}$$

This shows that in time series it is the asymmetry of the interaction between subsystems that leads to a significant lower bound to the irreversibility in terms of information.

For the basic linear signal-response model (BLRM discussed in [Auc+17]), in the limit of small observational time τ the backward transfer entropy converges to the causal influence. Also in the BLRM, we find that the causal influence rate converges to the Horowitz-Esposito[HE14] information flow.

A key quantity here is the observational time τ, and the detection of the irreversibility of processes from real (finite) time-series data is based on a fine-tuning of this parameter. We discuss this point with a biological model of receptor-ligand systems, where the entropy production measures the robustness of signaling.

Our motivation is a future application of the stochastic thermodynamics framework to the analysis of time series data in biology and finance. Here we will already consider its use in receptor-ligand systems, and in chapter 6 we will discuss its use in the study of the response to light perturbations in the circadian clock network.

5.2 Bivariate time series stochastic thermodynamics

5.2.1 Introduction to stochastic thermodynamics

Entropy production in heat baths

Let us consider an ensemble of trajectories generated by a Markovian (memoryless) continuous-time stochastic process composed of two interacting variables x and y subject to Brownian noise dW. The stochastic differential equations (SDEs)

describing such kind of processes can be written in the Ito interpretation[Shr04] as:

$$\begin{cases} dx = g_x(x,y)dt + \sqrt{D_x(x,y)}\ dW_x \\ dy = g_y(x,y)dt + \sqrt{D_y(x,y)}\ dW_y \end{cases} \tag{5.2}$$

where $D_x(x,y)$ and $D_y(x,y)$ are diffusion coefficients whose (x,y) dependence takes into account the case of multiplicative noise. Brownian motion is characterized by $\langle dW_i(t) dW_j(t') \rangle = \delta_{ij}\delta_{tt'}dt$. The dynamics in (5.2) is bipartite, that means conditionally independent in updating: $p(x_{t+dt}, y_{t+dt}|x_t, y_t) = p(x_{t+dt}|x_t, y_t) \cdot p(y_{t+dt}|x_t, y_t)$.

The bipartite structure of (5.2) is fundamental in stochastic thermodynamics, because it allows the identification[Sek98] and additive separation[HE14] of the entropy production of the thermal baths in separate contact with x and y subsystems, $ds_b = ds_b^x + ds_b^y$. These are given by the **detailed balance relation**[IS13; Cro99]:

$$ds_b^x = \ln \frac{p(x_{t+dt}|x_t, y_t)}{p(x_{t+dt}|\widetilde{x}_t, y_t)}, \tag{5.3}$$

where x_{t+dt} is defined as the event of variable x assuming value x_t at time $t + dt$, and similarly $\widetilde{x}_t$ is the event of variable x assuming value x_{t+dt} at time t. An analogous expression to (5.3) holds for subsystem y. Time-integrals of the updating probabilities $p(x_{t+dt}|x_t, y_t)$ and $p(x_{t+dt}|\widetilde{x}_t, y_t)$ can be written in terms of the SDE (5.2) using Onsager-Machlup action functionals[RH16; TC07].

Stochastic thermodynamics quantities are defined in single realizations of the probabilistic dynamics, in relation to the ensemble distribution[Sei12; Sei05]. As an example, the stochastic joint entropy is $s_{xy} = -\ln p_t(x_t, y_t)$, and its thermal (ensemble) average is the macroscopic entropy $S_{xy} = \langle s_{xy} \rangle_{p_t(x_t,y_t)}$. The explicit time dependence of p_t describes the ensemble dynamics, that in stationary processes is a relaxation to steady-state. A SDE system like (5.2) can be transformed into an equivalent partial differential equation in terms of probability currents[Ris96], that is the Fokker-Planck equation $\partial_t p_t(x,y) = -\partial_x J_x(x,y,t) - \partial_y J_y(x,y,t)$.

Feedbacks and information

The stochastic entropy of subsystem x *unaware* of the other subsystem y is $s_x = -\ln p_t(x_t)$, and its time variation is $ds_x = \ln \frac{p_t(x_t)}{p_{t+dt}(x_{t+dt})}$. The entropy production of subsystem x with its heat bath is $ds_{x+b} = ds_x + ds_b^x$, and its thermal average $\langle ds_{x+b} \rangle$ can be negative due to the interaction with y, in apparent violation of the thermodynamics II Law. This is the case of Maxwell's demon strategies[SU12; SU10], where information on x gained by the measuring device y is exploited to

exert a feedback force and extract work from x, like in the feedback cooling of a Brownian particle[HS14; RH16]. Integral fluctuation theorems[Par+15; IS13] (IFTs) provide lower bounds on the subsystems' macroscopic entropy production in terms of information-theoretic measures[CT12]. The stochastic mutual information is defined as $I_t^{st} \equiv \ln\left(\frac{p_t(x_t,y_t)}{p_t(x_t)p_t(y_t)}\right)$, where "$st$" stands for stochastic. Its time derivative can be separated into contributions corresponding to single trajectory movements and ensemble probability currents in the two directions, $d_t I_t^{st} = i_t^x + i_t^y$. The thermal average $I_{x\to y}(t) \equiv \langle i_t^y \rangle$ is the Horowitz-Esposito information flow[HE14; HS14]. At steady-state it takes the form:

$$I_{x\to y} = \int\int dxdy\ J_y(x,y)\frac{\partial \ln p(x|y)}{\partial y}. \tag{5.4}$$

A recent formulation[RH16] bounds the average work extracted at steady state with $\frac{\langle dW_{ext}\rangle}{T} = -\langle ds_b^x\rangle \leq dtI_{x\to y}$, where T is the temperature. This inequality is recovered with a different formulation in terms of transfer entropies[Ito16]: $\frac{\langle dW_{ext}\rangle}{T} \leq T_{x\to y}(dt) - T_{x\to y}(-dt) = dtI_{x\to y}$. Forward and backward stochastic transfer entropy[Spi+16] are respectively defined as $T_{x\to y}^{st}(dt) = \ln\left(\frac{p(y_{t+dt}|x_t,y_t)}{p(y_{t+dt}|y_t)}\right)$, and $T_{x\to y}^{st}(-dt) = \ln\left(\frac{p(y_t|x_{t+dt},y_{t+dt})}{p(y_t|y_{t+dt})}\right)$.

Irreversible entropy production

The total *irreversible* entropy production[HE14; HS14] of the joint system and thermal baths is:

$$ds_i^{xy} = ds_{xy} + ds_b^x + ds_b^y, \tag{5.5}$$

where $ds_{xy} = \ln\frac{p_t(x_t,y_t)}{p_{t+dt}(x_{t+dt},y_{t+dt})}$. If the ensemble is at steady state $p_{t+dt}(x_{t+dt},y_{t+dt}) = p_t(x_{t+dt},y_{t+dt}) = p(\widetilde{x}_t,\widetilde{y}_t)$. If we further assume that diffusion coefficients in (5.2) are nonzero constants, and this is the case of Langevin systems[Sek98] where these are proportional to the temperature, then the conditional probability $p(x_{t+dt}|\widetilde{x}_t,y_t)$ is equivalent to $p(x_{t+dt}|\widetilde{x}_t,y(t)=y_{t+dt}) = p(x_{t+dt}|\widetilde{x}_t,\widetilde{y}_t)$ under the integral sign[IS13]. Then the **irreversible entropy production** (5.5) takes the form:

$$ds_i^{xy} = \ln\left(\frac{p(x_t,y_t,x_{t+dt},y_{t+dt})}{p(\widetilde{x}_t,\widetilde{y}_t,x_{t+dt},y_{t+dt})}\right). \tag{5.6}$$

Equation (5.6) shows the connection between entropy production and irreversibility of trajectories. The thermal average has the form of a Kullback-Leibler divergence[CT12; RP12] and satisfy $dS_i^{xy} \equiv \langle ds_i^{xy}\rangle \geq 0$. The irreversible entropy production dS_i^{xy} is strictly positive when the interaction between subsystems is non-conservative[Che+06]. Our main interest is the stationary dissipation due to asymmetric interactions between subsystems.

5.2.2 Time series irreversibility measures

Setting and definition of causal representations

Let us assume that we are able to measure the state of the system (x, y) at a frequency $\frac{1}{\tau}$, thus obtaining time series. The finite observational time $\tau > 0$ makes the updating probability ***not bipartite***: $p(x_{t+\tau}, y_{t+\tau}|x_t, y_t) = p(x_{t+\tau}|x_t, y_t, y_{t+\tau}) \cdot p(y_{t+\tau}|x_t, y_t)$. Therefore a clear identification of thermodynamics quantities in time series is not possible. Let us take the Markovian SDE system (5.2) as the underlying process, and let us further assume stationarity. Then the statistical properties of time series obtained from a time discretization can be represented in the form of Bayesian networks, where links correspond to the way in which the joint probability density $p(x_t, y_t, x_{t+\tau}, y_{t+\tau})$ of states at the two instants t and $t + \tau$ is factorized. Still, there are multiple ways of such factorization. We say that a Bayesian network is a ***causal representation*** of the dynamics if conditional probabilities are expressed in a way that variables at time $t + \tau$ depend on variables at the same time instant or on variables at the previous time instant t (and not vice-versa), and that the dependence structure is done in order to minimize the total number of conditions on the probabilities. This corresponds to a minimization of the number of links in the Bayesian network describing the dynamics with observational time τ.

We define the combination ζ_τ^{xy} as a couple of successive states of the joint system (x, y) separated by a time interval τ, $\zeta_\tau^{xy} \equiv (x(t) = x_t, y(t) = y_t, x(t + \tau) = x_{t+\tau}, y(t + \tau) = y_{t+\tau}) \equiv f_\tau^{xy}(x_t, y_t, x_{t+\tau}, y_{t+\tau}) \equiv (x_t, y_t, x_{t+\tau}, y_{t+\tau})$. We use the identity functional $f_\tau^{xy}(a, b, c, d) \equiv (x(t) = a, y(t) = b, x(t + \tau) = c, y(t + \tau) = d)$ for an unambiguous specification of the backward combination $\widetilde{\zeta_\tau^{xy}}$. This is defined as the time-reversed conjugate of the combination ζ_τ^{xy}, meaning the inverted couple of the same two successive states, $\widetilde{\zeta_\tau^{xy}} \equiv f_\tau^{xy}(x_{t+\tau}, y_{t+\tau}, x_t, y_t) \equiv (\widetilde{x_t}, \widetilde{y_t}, \widetilde{x_{t+\tau}}, \widetilde{y_{t+\tau}})$. We defined backward variables of the type $\widetilde{x_t}$ meaning $x(t) = x_{t+\tau}$, such correspondences being possible only when states at both times t and $t + \tau$ are given. The subsystems variables and backward variables are similarly defined as $\zeta_\tau^x = (x_t, x_{t+\tau})$, $\zeta_\tau^y = (y_t, y_{t+\tau})$, $\widetilde{\zeta_\tau^x} = (\widetilde{x_t}, \widetilde{x_{t+\tau}})$, and $\widetilde{\zeta_\tau^y} = (\widetilde{y_t}, \widetilde{y_{t+\tau}})$.

Definition of mapping irreversibility and the standard fluctuation theorem

A measure of **coarse grained entropy production** for time series can be defined replacing dt with the nonzero observational time τ in the general expression (5.5) obtaining:

$$\begin{aligned} \Delta s_i^{xy} = \Delta s_{xy} + \Delta s_m^x + \Delta s_m^y \equiv \\ \equiv \ln \frac{p(x_t,y_t)}{p(x_{t+\tau},y_{t+\tau})} + \ln \frac{p(x_{t+\tau}|x_t,y_t)}{p(\widetilde{x_{t+\tau}}|\widetilde{x_t},y_t)} + \ln \frac{p(y_{t+\tau}|y_t,x_t)}{p(\widetilde{y_{t+\tau}}|\widetilde{y_t},x_t)}, \end{aligned} \tag{5.7}$$

where we assumed stationarity, $p_t = p$. By definition Δs_i^{xy} converges to the physical entropy production in the limit $\tau \to 0$, and it is a lower bound to it[GM+08]. Importantly, such coarse grained entropy production cannot have the form of an irreversibility measure like (5.6) because $p(y_{t+\tau}|y_t, x_t, x_{t+\tau}) \neq p(y_{t+\tau}|y_t, x_t)$. With "irreversibility form" we mean that its thermal average is a Kullback-Leibler divergence measuring the distinguishability between forward and time-reverse paths. Therefore we decided to keep the widely accepted time series irreversibility definition given in [RP12], in its form for stationary Markovian systems.

For the study of fluctuations, we define the **stochastic mapping irreversibility** with observational time τ of the joint system (x, y) as:

$$\varphi_\tau^{xy} = \ln \left(\frac{p(\zeta_\tau^{xy})}{p(\widetilde{\zeta_\tau^{xy}})} \right). \tag{5.8}$$

The thermal average $\Phi_\tau^{xy} \equiv \langle \varphi_\tau^{xy} \rangle_{p(\zeta_\tau^{xy})}$ describes the statistical properties of a single transition over an interval τ. However, since the underlying dynamics (5.2) is Markovian, it describes the irreversibility of arbitrary long time series. Importantly, φ_τ^{xy} does not generally converge to the total physical entropy production (5.5) in the limit $\tau \to 0$ because of the non-bipartite structure of $p(x_{t+\tau}, y_{t+\tau}|x_t, y_t)$ for any $\tau > 0$. It does converge anyway in most physical situations where a bipartite underlying dynamics like (5.2) has constant and strictly positive diffusion coefficients, and this is the case of Langevin systems. This is because the Brownian increments dW_x and dW_y are dominating the dynamics for small intervals τ, then the estimate of $f_y(x_{t+t'}, y_{t+t'})$ (with $0 \leq t' \leq \tau$) based on (x_t, y_t) is improved with the knowledge of $x_{t+\tau}$ just by a term of order $\partial_x f_y(x, y) \cdot W_x(\tau) \sim \sqrt{\tau}$, where we assumed a smooth $f_y(x, y)$. Therefore in the limit $\tau \to 0$ it is almost surely $p(y_{t+\tau}|x_t, y_t, x_{t+\tau}) \to p(y_{t+\tau}|x_t, y_t)$ and $\varphi_\tau^{xy} \to \Delta s_i^{xy} \to ds_i^{xy}$, see Appendix D.

We are interested in the time-reversal asymmetry of time series from even more general models or data where no identification of thermodynamic quantities is required.

The stochastic mapping irreversibility satisfies the integral fluctuation theorem[Sei12], i.e.:

$$\left\langle e^{-\varphi_\tau^{xy}} \right\rangle_{p(\zeta_\tau^{xy})} = \int_\Omega d\zeta_\tau^{xy} p(\widetilde{\zeta_\tau^{xy}}) = 1, \tag{5.9}$$

where $d\zeta_\tau^{xy} = dx_t dy_t dx_{t+\tau} dy_{t+\tau}$, $\widetilde{dx_t} = dx_{t+\tau}$, and Ω is the whole space of the combination ζ_τ^{xy}. From the convexity of the exponential function it follows that the entropy production Φ_τ^{xy}, which is the ensemble average of the stochastic entropy production φ_τ^{xy}, is non-negative. This is the standard thermodynamics II Law inequality for the joint system (x, y) time series:

$$\Phi_\tau^{xy} = \langle \varphi_\tau^{xy} \rangle_{p(\zeta_\tau^{xy})} \geq 0. \tag{5.10}$$

Similarly, we define the stochastic mapping irreversibility for the two subsystems as $\varphi_\tau^x \equiv \ln\left(\frac{p(\zeta_\tau^x)}{p(\widetilde{\zeta_\tau^x})}\right)$ and $\varphi_\tau^y \equiv \ln\left(\frac{p(\zeta_\tau^y)}{p(\widetilde{\zeta_\tau^y})}\right)$. Their ensemble averages are respectively denoted $\Phi_\tau^x \geq 0$ and $\Phi_\tau^y \geq 0$, and they also satisfy the standard II Law. Importantly, although bivariate time series derived from the joint process (5.2) are Markovian, the one-dimensional subsystems time series are generally not. This is because subsystems trajectories are a coarse grained representation of the full dynamics and in order to reproduce the statistical properties of trajectories, a non-Markovian dynamics has to be assumed. Therefore Φ_τ^x and Φ_τ^y are generally different from the irreversibility calculated on a whole time series. The mapping irreversibility Φ_τ^x describes the statistical properties of the whole time series only if it is Markovian. This is surely the case if x is not influenced by y in (5.2), $\partial_y g_x(x, y) = \partial_y D_x(x, y) = 0$, and motivated our study of signal-response models[Auc+17].

We define the **conditional mapping irreversibility** of y given x as the difference between the mapping irreversibility of the joint system (x, y) and the mapping irreversibility of system x alone[CS19]:

$$\begin{aligned} \Phi_\tau^{y|x} &= \Phi_\tau^{xy} - \Phi_\tau^x = \\ &= \left\langle \ln\left(\frac{p(y_t, y_{t+\tau}|x_t, x_{t+\tau})}{p(\widetilde{y_t}, \widetilde{y_{t+\tau}}|\widetilde{x_t}, \widetilde{x_{t+\tau}})}\right) \right\rangle_{p(\zeta_\tau^{xy})}. \end{aligned} \tag{5.11}$$

We note that conditioning on the x combination ζ_τ^x introduces an anti-causal form, meaning that (5.11) includes conditions on future states for the conditional probabilities of past states. The anti-causal form arises necessarily as a consequence of the non-bipartite structure of time series.

In the general case where the evolution of each variable is influenced by the other variable (Eq.5.2), we have a complete causal representation resulting from the dynamics (Fig.5.1), meaning that all edges are present in the Bayesian network. In this case we were not able to provide a more accurate characterization of the

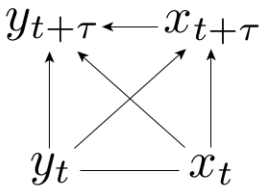


Fig. 5.1: Complete causal representation. The arrows represent the way we decompose the joint probability density. In the complete case we have $p(\zeta_\tau^{xy}) = p(x_t, y_t) \cdot p(x_{t+\tau}|x_t, y_t) \cdot p(y_{t+\tau}|x_t, y_t, x_{t+\tau})$.

mapping irreversibility Φ_τ^{xy} than the one given by the standard thermodynamics II Law (Eq.5.10). We aim to relate the irreversibility of the joint time series to the information flow between subsystems variables over time. We argue that more informative fluctuation theorems arise as a consequence of missing edges in the causal representation of the dynamics in terms of Bayesian networks. In the bivariate case there is only one class of continuous-time models for which informative fluctuation theorems for causal representations can be written: the signal-response models.

The mapping irreversibility density

Let us use an equivalent representation of the mapping irreversibility in terms of backward probabilities[IS16] defined as $p_B(\zeta_\tau^{xy}) = p(x(t) = x_t, y(t) = y_t, x(t-\tau) = x_{t+\tau}, y(t-\tau) = y_{t+\tau})$. For stationary processes it holds $p_B(\zeta_\tau^{xy}) = p(\widetilde{\zeta_\tau^{xy}})$ and $\varphi_\tau^{xy} = \ln\Big(\frac{p(\zeta_\tau^{xy})}{p_B(\zeta_\tau^{xy})}\Big)$. We introduce here the **mapping irreversibility density** (with observational time τ) for stationary processes as:

$$\begin{aligned}
\psi(x_t, y_t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx_{t+\tau}dy_{t+\tau}p(\zeta_\tau^{xy})\varphi_\tau^{xy} = \\
= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx_{t+\tau}dy_{t+\tau}p(\zeta_\tau^{xy}) \ln\left(\frac{p(\zeta_\tau^{xy})}{p_B(\zeta_\tau^{xy})}\right) \\
= p(x_t, y_t)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx_{t+\tau}dy_{t+\tau}p(x_{t+\tau}, y_{t+\tau}|x_t, y_t) * \\
* \ln\left(\frac{p(x(t+\tau)=x_{t+\tau}, y(t+\tau)=y_{t+\tau}|x(t)=x_t, y(t)=y_t)}{p(x(t-\tau)=x_{t+\tau}, y(t-\tau)=y_{t+\tau}|x(t)=x_t, y(t)=y_t)}\right).
\end{aligned} \tag{5.12}$$

The mapping irreversibility density $\psi(x_t, y_t)$ tells us which situations (x_t, y_t) contribute more to the time series irreversibility of the macroscopic process. $\psi(x_t, y_t)$ is proportional to the distance (precisely to the Kullback–Leibler divergence[CT12]) of the distribution of future states $p(x_{t+\tau}, y_{t+\tau}|x_t, y_t)$ to the distribution of past states $p(x_{t-\tau}, y_{t-\tau}|x_t, y_t)$ of the same condition (x_t, y_t).

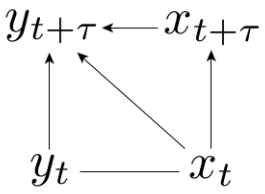


Fig. 5.2: Causal representation of signal-response models. The joint probability density is decomposed into $p(\zeta_\tau^{xy}) = p(x_t, y_t) \cdot p(x_{t+\tau}|x_t) \cdot p(y_{t+\tau}|x_t, y_t, x_{t+\tau})$.

5.2.3 The fluctuation theorem for signal-response models

If the system (x, y) is such that the variable y does not influence the dynamics of the variable x, then we are dealing with signal-response models (Fig.5.2). The stochastic differential equation for signal-response models is written in the Ito representation[Shr04] as:

$$\begin{cases} dx = g_x(x)dt + \sqrt{D_x(x)}\ dW_x \\ dy = g_y(x,y)dt + \sqrt{D_y(x,y)}\ dW_y \end{cases} \tag{5.13}$$

The absence of feedback is written in $\frac{\partial g_x}{\partial y} = \frac{\partial D_x}{\partial y} = 0$. As a consequence the conditional probability satisfies $p(y_t|x_t, x_{t+\tau}) = p(y_t|x_t)$, and the corresponding causal representation is incomplete, see the Bayesian network in Fig.5.2.

For signal-response models we can provide a lower bound on the entropy production that is more informative than Eq.5.10, and that involves the backward transfer entropy $T_{y\to x}(-\tau)$. The backward transfer entropy[Ito16] is a measure of discrete information flow towards past, and is here defined as the standard transfer entropy for the ensemble of time-reversed combinations ζ_τ^{xy}. The stochastic counterpart as a function of $\zeta_\tau^{xy} \setminus y_t$ is defined as:

$$T^{st}_{y\to x}(-\tau) = \ln\left(\frac{p(x_t|y_{t+\tau}, x_{t+\tau})}{p(x_t|x_{t+\tau})}\right), \tag{5.14}$$

where st stands for stochastic.

Then by definition $T_{y\to x}(-\tau) = \left\langle T^{st}_{y\to x}(-\tau)\right\rangle_{p(\zeta_\tau^{xy})\setminus y_t}$. We keep the same symbol $T_{y\to x}$ as the standard transfer entropy because in stationary processes the backward transfer entropy is the standard transfer entropy (calculated on forward trajectories) for negative shifts $-\tau$.

The **fluctuation theorem for signal-response models** is written:

$$\left\langle e^{-\varphi_\tau^{xy}+\varphi_\tau^x+T^{st}_{y\to x}(-\tau)}\right\rangle_{p(\zeta_\tau^{xy})} = \\ = \int_\Omega d\zeta_\tau^{xy} p(\widetilde{y_{t+\tau}}|\widetilde{x}_t,\widetilde{y}_t,\widetilde{x_{t+\tau}})p(x_t,x_{t+\tau},y_{t+\tau}) = 1, \tag{5.15}$$

where we used the signal-response property of no feedback $p(\widetilde{y}_t|\widetilde{x}_t,\widetilde{x_{t+\tau}}) = p(\widetilde{y}_t|\widetilde{x}_t)$, the correspondence $dy_t = dy_{t+\tau}$, and the normalization $\int_{-\infty}^{\infty} d\widetilde{y_{t+\tau}} p(\widetilde{y_{t+\tau}}|\widetilde{x}_t,\widetilde{y}_t,\widetilde{x_{t+\tau}}) = 1$.

From the convexity of the exponential it follows the **II Law-like inequality for signal-response models** (Eq.5.1):

$$\Phi_\tau^{y|x} = \Phi_\tau^{xy} - \Phi_\tau^x \geq T_{y\to x}(-\tau),$$

and this is our main result. Let us note that in the limit of $\tau \to 0$ and constant nonzero diffusion coefficients, $\Phi_\tau^{y|x}$ converges to the physical entropy production $\langle ds_m^y \rangle$, and the inequality (5.1) converges to a special case of a previous result on bipartite systems[Ito16]. Note that Φ_τ^x is equivalent to the original time series irreversibility[RP12] because the x time series is Markovian in the absence of feedback.

In causal representations of correlated stationary processes the factorization of $p(x_t, y_t)$ is unnecessary, and only the transition probability $p(x_{t+\tau}, y_{t+\tau}|x_t, y_t)$ makes the difference, and we don't specify the direction of the x_t-y_t arrow in Fig.5.1-5.2. The importance of the causal representation is seen here because we could have decomposed the transition probability as well into the non-causal decomposition $p(x_{t+\tau}, y_{t+\tau}|x_t, y_t) = p(y_{t+\tau}|x_t, y_t) \cdot p(x_{t+\tau}|x_t, y_t, y_{t+\tau})$, but this does not lead to the fluctuation theorem (5.15).

5.3 Applications

5.3.1 The basic linear response model

We study the II law for signal-response models (Eq.5.1) in the basic linear response model (BLRM), whose information processing properties for the forward trajectories have already been discussed in chapter 4 (and in [Auc+17]). The BLRM is composed

of a fluctuating signal x described by the Ornstein-Uhlenbeck process[UO30; Gil96a], and a dynamic linear response y to this signal:

$$\begin{cases} dx = -\frac{x}{t_{rel}}dt + \sqrt{D}\ dW \\ \frac{dy}{dt} = \alpha x - \beta y \end{cases} \tag{5.16}$$

The response y is considered in the limit of weak coupling with the thermal bath $D_y \to 0$, while the signal is attached to the source of noise, $D_x = D > 0$.

This model allows analytical representations for the mapping irreversibility Φ_τ^{xy} (calculated in Appendix A) and the backward transfer entropy $T_{y\to x}(-\tau)$ (calculated in Appendix B). We find that, once the observational time τ is specified, Φ_τ^{xy} and $T_{y\to x}(-\tau)$ are both functions of just the two parameters t_{rel} and β, which describe respectively the time scale of the fluctuations of the signal and the time scale of the response to a deterministic input.

Since the signal is a time-symmetric (reversible) process, $\Phi_\tau^x = 0$, the backward transfer entropy $T_{y\to x}(-\tau)$ is the lower bound on the total entropy production Φ_τ^{xy} in the BLRM.

The plot in Fig.5.3 shows the mapping irreversibility Φ_τ^{xy} and the backward transfer entropy $T_{y\to x}(-\tau)$ as a function of the observational time τ. In the limit of small τ, the entropy production diverges because of the deterministic nature of the response dynamics (the standard deviation on the determination of the velocity $\frac{dy}{dt}$ due to instantaneous movements of the signal vanishes as $\alpha\sqrt{D}\sqrt{dt} \to 0$). The backward transfer entropy $T_{y\to x}(-\tau)$ instead vanishes for $\tau \to 0$ because the Brownian motion has nonzero quadratic variation[Shr04] and is the dominating term in the signal dynamics for small time intervals. In the limit of large observational time intervals $\tau \to \infty$ the entropy production is asymptotically double the backward transfer entropy, that is its lower bound given by the II law for signal-response models (Eq.5.1), $\frac{\Phi_\tau^{xy}}{T_{y\to x}(-\tau)} \to 2$. Surprisingly, this limit of 2 is valid for any choice of the parameters in the BLRM.

Interestingly, for small observational time $\tau \to 0$, the causal influence of the signal on the evolution of the response (defined in chapter 4 and in [Auc+17]) converges to the backward transfer entropy of the response on the past of the signal $C_{x\to y}(\tau) \to T_{y\to x}(-\tau)$, and they both vanish with $\tau\beta$. Also, the causal influence rate defined as $\lim_{\tau\to 0} \frac{C_{x\to y}(\tau)}{\tau}$ converges to the Horowitz-Esposito[HE14] information flow $I^{x\to y}$ (details in Appendix C).

For large observational time $\tau \to \infty$ instead the causal influence converges to the standard (forward) transfer entropy $C_{x\to y}(\tau) \to T_{y\to x}(\tau)$. Also in this limit $\tau \to \infty$,

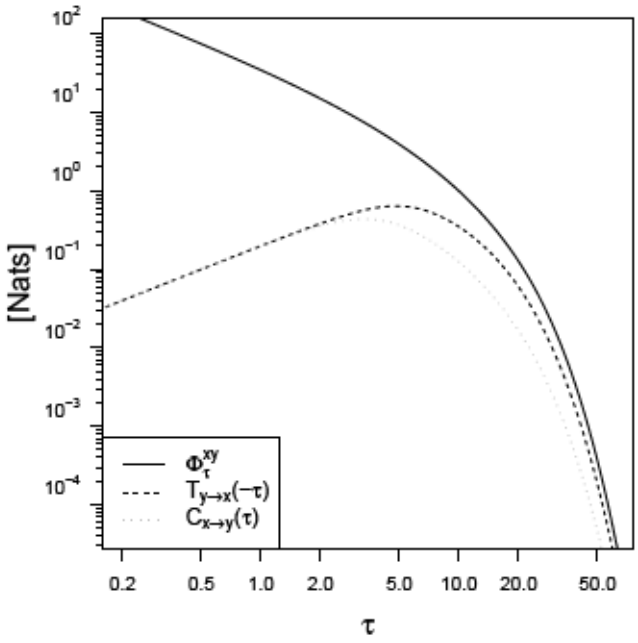


Fig. 5.3: Mapping irreversibility Φ_τ^{xy}, backward transfer entropy $T_{y\to x}(-\tau)$ and causal influence $C_{x\to y}(\tau)$ in the the BLRM as a function of the observational time interval τ. The parameters are $\beta = 0.2$ and $t_{rel} = 10$. All graphs are produced using R[R C14].

the causal influence is an eighth of the entropy production $\frac{\Phi_\tau^{xy}}{C_{x\to y}(\tau)} \to 8$ for any choice of the parameters in the BLRM.

Let us now consider the mapping irreversibility density $\psi(x_t, y_t)$ in the BLRM for small and large observational time τ. In Fig.5.4 we choose a τ smaller than the characteristic response time $\frac{1}{\beta}$ and also smaller than the characteristic time of fluctuations t_{rel}. In the limit $\tau \to 0$ the signal dynamics is dominated by noise and the entropy production is mainly given by movements of the response y. The two spots correspond to the points where the product of the density $p(x_t, y_t)$ times the absolute value of the instant velocity $\dot{y}$ is larger. For longer intervals $\tau \quad \frac{1}{\beta}$ (that is the case of Fig.5.5) the history of the signal becomes relevant because it determined the present value of the response, therefore the irreversibility density is also distributed on those points of the diagonal (corresponding to roughly $\dot{y} = 0$) where the absolute value of the response y is big enough. Also as a consequence, in this regime the backward transfer entropy is a meaningful lower bound on the entropy production, that is Eq.5.1.

Addition of intrinsic noise in the BLRM

We choose to work the basic ideas in a model (the BLRM) with no intrinsic noise in the response, meaning a deterministic channel, and the effective noise source in the information $I(x_t, y_t)$ is given by instances of the signal at previous times. Let us now briefly discuss what it would change if we consider the BLRM with additional noise of intensity D_y in the response. The dynamics is then written:

$$\begin{cases} dx = -\frac{x}{t_{rel}}dt + \sqrt{D_x}\, dW^{(x)} \\ dy = (\alpha x - \beta y)\, dt + \sqrt{D_y}\, dW^{(y)} \end{cases} \tag{5.17}$$

The formal solution becomes:

$$y_{t+\tau} = y_t e^{-\beta\tau} + \alpha \int_0^\tau dt'\, x_{t+t'} e^{-\beta(\tau - t')} + \sqrt{D_y} \int_0^\tau dW_{t+t'}^{(y)} e^{-\beta(\tau - t')}. \tag{5.18}$$

The additional noise affects only the dynamics of the y and doesn't affect the expectation of $\langle x_t y_{t+\tau} \rangle$, therefore it reduces the mutual information $I(x_t, y_{t+\tau})$ only through a change of the response variance σ_y^2. Let us recall that the nonzero quadratic variation of Brownian motion makes the second order derivatives of y functions not vanishing in the Ito interpretation od SDEs. So we calculate:

$$0 = d\langle y^2 \rangle = 2\langle y dy \rangle + \langle (dy)^2 \rangle = 2\langle y dy \rangle + D_y dt. \tag{5.19}$$

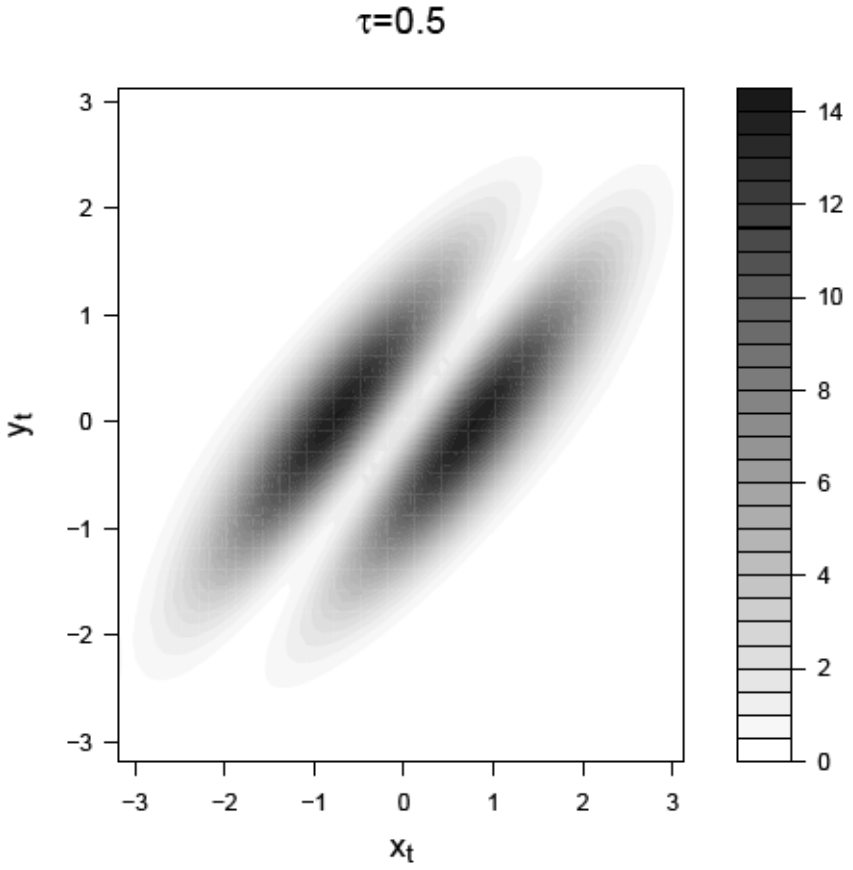


Fig. 5.4: Mapping irreversibility density $\psi(x_t, y_t)$ for the BLRM at $\tau = 0.5 < \frac{1}{\beta} < t_{rel}$. The parameters are $\beta = 0.2$ and $t_{rel} = 10$. Both $\psi(x_t, y_t)$ and the space (x, y) are expressed in units of standard deviations.

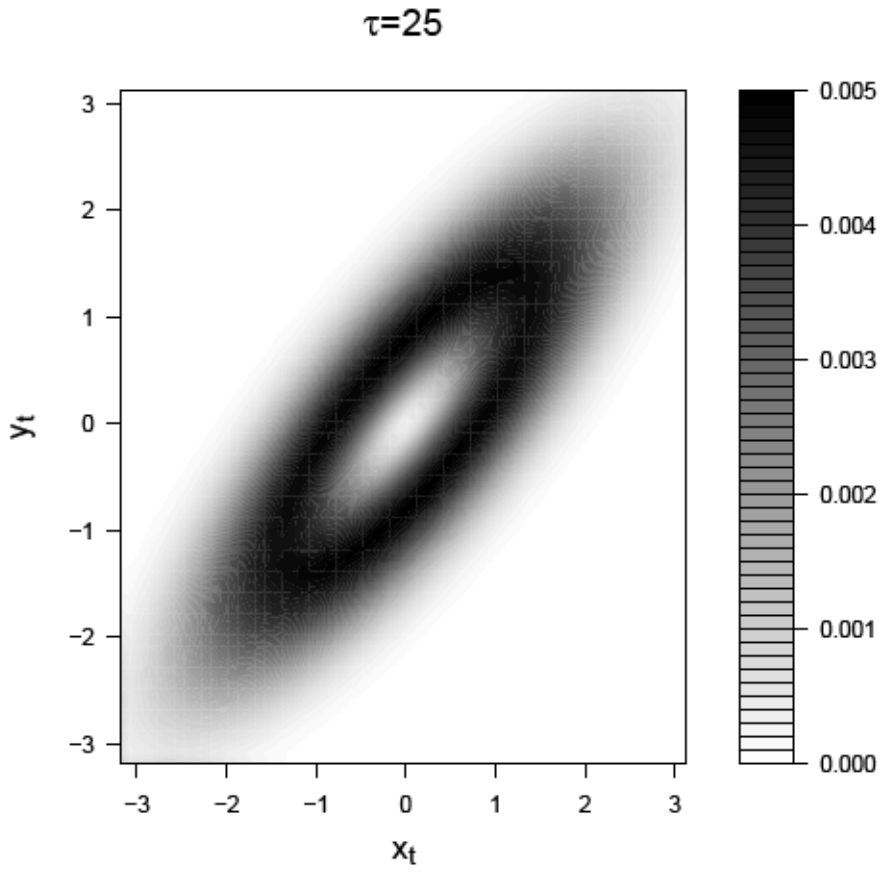


Fig. 5.5: Mapping irreversibility density $\psi(x_t, y_t)$ for the BLRM at $\tau = 25 > t_{rel} > \frac{1}{\beta}$. The parameters are $\beta = 0.2$ and $t_{rel} = 10$. Both $\psi(x_t, y_t)$ and the space (x, y) are expressed in units of standard deviations.

Then adding the additional term to the response variance we obtain (see Chapter 4, pg.64):

$$\sigma_y^2 = \frac{\alpha^2 \sigma_x^2}{\beta(\beta + \frac{1}{t_{rel}})} + \frac{D_y}{2\beta}. \tag{5.20}$$

The additional term in the conditional variance $\sigma^2_{y_{t+\tau}|x_t,y_t}$ is calculated as:

$$\begin{aligned}\sigma^2_{y_{t+\tau}|x_t,y_t}(D_y) = \sigma^2_{y_{t+\tau}|x_t,y_t}|_{D_y=0} + D_y e^{-2\beta\tau} \left\langle \int_0^\tau \int_0^\tau dW^{(y)}_{t+t'} dW^{(y)}_{t+t''} e^{\beta(t'+t'')} \right\rangle = \\ = \sigma^2_{y_{t+\tau}|x_t,y_t}|_{D_y=0} + \frac{D_y}{2\beta}(1 - e^{-2\beta\tau}),\end{aligned} \tag{5.21}$$

and similarly $\sigma^2_{y_{t+\tau}|x_t,y_t,x_{t+\tau}} = \sigma^2_{y_{t+\tau}|x_t,y_t,x_{t+\tau}}|_{D_y=0} + \frac{D_y}{2\beta}(1 - e^{-2\beta\tau})$.

Now it is important to note how the additional noise affects the causal influence and the irreversibility measures. For small $D_y << \beta\sigma_y^2|_{D_y=0}$, the mutual information does not qualitatively changes because $\sigma^2_{y_{t+\tau}|x_t}|_{D_y=0}$ is finite, and the same we can say for the Redundancy $R(\tau)$. Then the causal influence $C_{x\to y}$ changes continuously when the additional noise is introduced, and it progressively decreases with D_y.

The situation is different for the irreversibility measure. The mapping irreversibility Φ^{xy}_τ diverges in the BLRM (that is system (5.17) with $D_y = 0$) because, given x_t and y_t, the estimate of $y_{t+\tau}$ differs from the estimate of $y_{t-\tau}$ by a term $\simeq \tau\ 2\alpha x_t$, whose variance is proportional to $\sigma_x \tau^2$, while the uncertainty on these estimates is described by the variance $\sigma_{y_{t+\tau}|x_t,y_t} \sim \tau^3$. When the additional noise source in y is introduced, $D_y > 0$, the variance goes with $\sigma_{y_{t+\tau}|x_t,y_t} \sim \tau$, as it is always the case for Brownian motion, and the mapping irreversibility vanishes for $\tau \to 0$.

5.3.2 Receptor-ligand systems

The Receptor-Ligand interaction is the fundamental mechanism of molecular recognition in biology and is a recurring motif in signaling pathways[Kli+16; Kho06]. The fraction of activated receptors is part of the cell's representation of the outside world, it is the cell's estimate on the concentration of ligands in the environment, upon which it bases its protein expression and response to external stress.

If we could experimentally keep the concentration of ligands fixed we would still get a fluctuating number of activated receptors due to the intrinsic stochasticity of the macroscopic description of chemical reactions. Recent studies allowed a theoretical understanding of the origins of the macroscopic "noise" (i.e. the output variance in the conditional probability distributions), and also raised questions about the optimality of the input distributions in terms of information transmission[BS05; Tka+08b; Cri+18; WK11].

Here we consider the dynamical aspects of information processing in receptor-ligand systems[DTW12; Nem12], where the response is integrated over time. If the perturbation of the receptor-ligand binding on the concentration of free ligands is negligible, that is in the limit of high ligand concentration, receptor-ligand systems can be modeled as nonlinear signal-response models[DTW14]. We write our model of receptor-ligand systems in the Ito representation[Shr04] as:

$$\begin{cases} dx = -(x-1)dt + x\, dW_x \\ dy = k_{on}(1-y)\frac{x^h}{1+x^h}dt - k_{off}ydt + y(1-y)dW_y \end{cases} \tag{5.22}$$

The fluctuations of the ligand concentration x are described by a mean-reverting geometric Brownian motion, with an average $\langle x \rangle = 1$ in arbitrary units. The response, that is the fraction of activated receptors y, is driven by a Hill-type interaction with the signal with cooperativity coefficient h, and chemical bound/unbound rates k_{on} and k_{off}. For simplicity, the dynamic range of the response is set to be coincident with the mean value of the ligand concentration, that means to choose a Hill constant $K = \langle x \rangle = 1$. The form of the y noise is set by the biological constraint $0 < y < 1$. For simplicity, we choose a cooperativity coefficient of $h = 2$, that is the lower order of sigmoidal functions. A sample of the activated receptors fraction and ligand concentration dynamics is plotted in Fig.5.6.

The mutual information between the concentration of ligands and the fraction of activated receptors in a cell is a natural choice for quantifying its sensory properties[Tka+09]. Here we argue that, in the case of signal-response models, the conditional entropy production is the relevant measure, because it quantifies how the dynamics of the signal produces irreversible transitions in the dynamics of the response, which is closely related to the concept of causation. Besides, our measure of causal influence[Auc+17] has yet not been generalized to the nonlinear case, while the entropy production has a consistent thermodynamical interpretation[Sei12].

We simulated the receptor-ligand model of Eq.5.22, and we evaluated numerically the mapping irreversibility Φ_τ^{xy} and the backward transfer entropy $T_{y\to x}(-\tau)$ using a multivariate Gaussian approximation for the conditional probabilities $p(x_{t+\tau}, y_{t+\tau}|x_t, y_t)$ (details in Appendix E). The II law for signal response models sets $\Phi_\tau^{xy} \geq T_{y\to x}(-\tau)$ and proves to be a useful tool for receptor-ligand systems, as it is seen if Fig.5.7. Note that the numerical estimation of the entropy production requires many more samples compared to the backward transfer entropy (see Appendix F): Φ_τ^{xy} depends on ζ_τ^{xy} (4 dimensions) while $T_{y\to x}(-\tau)$ depends on $\zeta_\tau^{xy} \setminus y_t$ (3 dimensions). In a real biological experimental setting the sampling process is expensive, and the backward transfer entropy is therefore a useful lower bound for the entropy production, and an interesting characterization of the system to be used when the number of samples is not large enough.

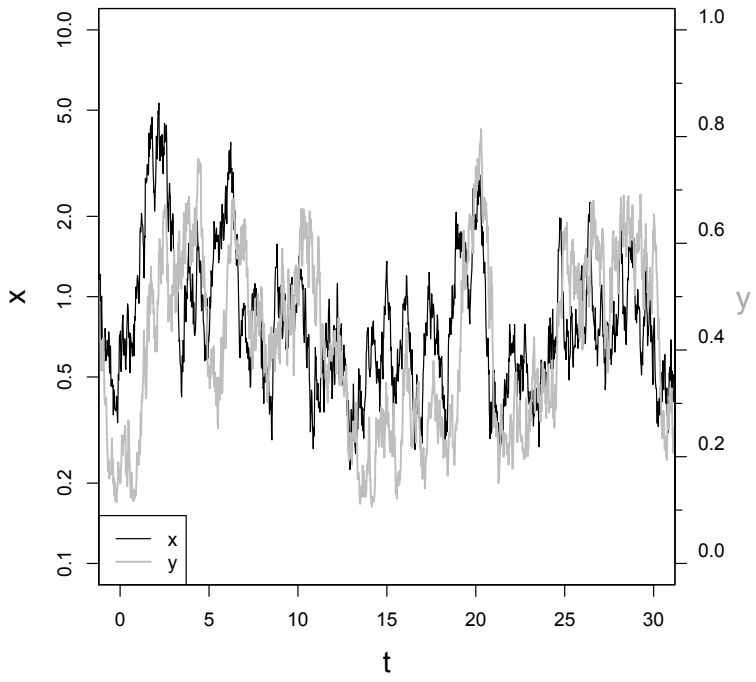


Fig. 5.6: Stochastic dynamics of the fraction of activated receptors (gray curve) and of the ligand concentration (black curve). The parameters are $k_{off} = 1$, $k_{on} = 2k_{off}$, $h = 2$, and $t_{rel} = 10$.

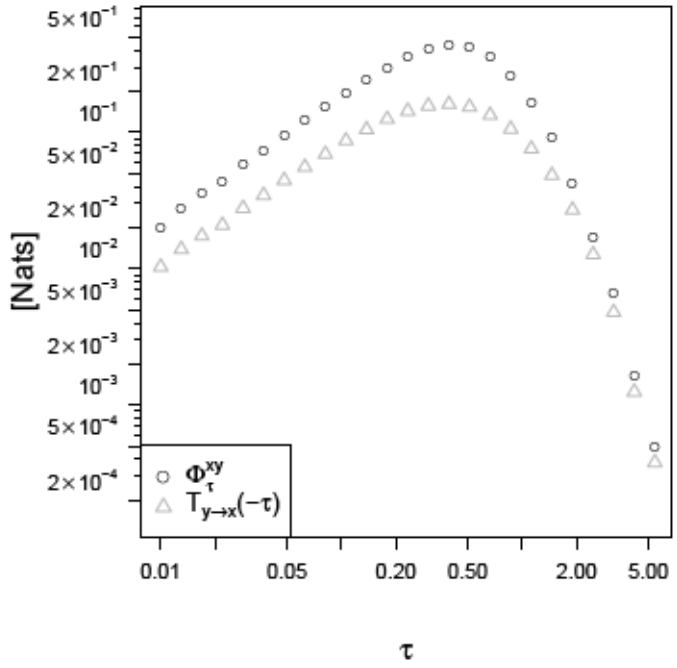


Fig. 5.7: Mapping irreversibility and backward transfer entropy in our model of receptor-ligand systems (Eq.5.22). The parameters are $k_{on} = 5$, $k_{off} = 1$, $h = 2$, and $t_{rel} = 10$.

The intrinsic noise of the response $y(1-y)dW_y$ is the dominant term in the response dynamics for small intervals τ. This makes both Φ_τ^{xy} and $T_{y\to x}(-\tau)$ vanish in the limit $\tau \to 0$. In the limit of large observational time τ, as it is also the case for the BLRM and in any stationary process, the entropy production for the corresponding time series Φ_τ^{xy} and all the information measures are vanishing, because the memory of the system is damped exponentially over time by the relaxation parameter k_{off} (β in the BLRM). Therefore in order to better detect the irreversibility of a process one must choose an appropriate observational time τ. In the receptor-ligand model of Eq.5.22 with parameters $k_{on} = 5$, $k_{off} = 1$, $h = 2$ and $t_{rel} = 10$ we see that the optimal observational time is around $\tau \approx 0.5$ (see Fig.5.7). Here for "optimal" we mean the observational time that corresponds to the highest mapping irreversibility Φ_τ^{xy}, but one might also be interested in inferring the entropy rate (that is $\frac{\Phi_\tau^{xy}}{\tau}$ in the limit $\tau \to 0$) looking at time series data with finite sampling interval τ. We do not treat this problem here.

5.4 Discussion

To put in perspective our work let us recall that the well-established integral fluctuation theorem for stochastic trajectories[Sei05] leads to a total irreversible entropy production with zero lower bound, which is the standard II Law of Thermodynamics. Our aim here was to characterize cases in which more informative lower bounds on the total entropy production can be provided. Ito-Sagawa[IS13] already showed that for Bayesian controlled systems (where a parameter can be varied to perform work) a general fluctuation theorem and the relative lower bound on entropy production is linked to the topology of the Bayesian network representation associated to the stochastic dynamics of the system. This connection seems to be even stronger in the case of uncontrolled systems that is the object of our study. We show in the bidimensional case of a pair of signal-response variables how a missing arrow in the Bayesian network describing the dynamics leads to a fluctuation theorem.

The detailed fluctuation theorem linking work dissipation and the irreversibility of trajectories in nonequilibrium transformations[Cro99; Jar00] holds in mechanical systems attached to heat reservoirs. We are interested here in the irreversibility of trajectories in more general models with asymmetric interactions, since this is mostly the case in biological systems or asset pricing models in quantitative finance. In those models there is no Hamiltonian description of work and heat, no microscopic reversibility, and the detailed fluctuation theorem is, properly, not a theorem but itself a definition of irreversibility.

We study time series resulting from a discretization with observational time τ of continuous stochastic processes. Importantly the underlying bipartite process appears, at limited time resolution, as a non-bipartite process. As a consequence there is no general convergence of the time series irreversibility to the physical entropy production except for special cases like Langevin systems. Our mapping irreversibility (5.8) is the Markovian approximation of the time series irreversibility definition given in [RP12]. While it is well defined for any stationary process, it describes the statistical properties of long time series only in the Markovian case.

For a general interacting dynamics like (5.2) we are not able to provide a more significant lower bound to the mapping irreversibility than the standard II law of thermodynamics (5.10). A more informative lower bound on the mapping irreversibility is found for signal-response models described by the absence of feedback, see (5.13). This sets the backward transfer entropy as a lower bound to the conditional entropy production, and describes the connection between the irreversibility of stochastic trajectories and the discrete information flow towards past between variables.

We restrict ourselves to the bivariate case here, but we conjecture that fluctuation theorems for multidimensional stochastic autonomous dynamics should arise in general as a consequence of missing arrows in the (non complete, see e.g. Fig.5.2) causal representation of the dynamics in terms of Bayesian networks.

In our opinion, a general relation connecting the incompleteness of the causal representation of the dynamics with information thermodynamics fluctuation theorems is still lacking.

Finally, let us note that exponential averages like our integral fluctuation theorem (5.15) are dominated by (exponentially) rare realizations[Jar06b], and the corresponding II Law inequalities like our (5.1) are often not very strict bounds. In the receptor-ligand model discussed in section II.B the backward transfer entropy lower bound is roughly one half of the mapping irreversibility, and this is also the case in the BLRM for large τ where the ratio converges exactly to $\frac{1}{2}$. This limitation is quite general, see for example the information thermodynamic bounds on signaling robustness given in [IS15].

We also introduced a discussion about the observational time τ in data analysis. In a biological model of receptor-ligand systems we showed that it has to be fine-tuned for a robust detection of the irreversibility of the process, which is related to the concept of causation[Auc+17] and therefore to the efficiency of biological coupling between signalling and response.

5.4.1 Appendix A: Mapping irreversibility in the BLRM

Let us consider an ensemble of stochastic trajectories generated with the BLRM (Eq.5.16). The mapping irreversibility Φ_τ^{xy} here is the Kullback-Leibler divergence[CT12] between the probability density $p(\zeta_\tau^{xy})$ of couples of successive states ζ_τ^{xy} separated by a time interval τ of the original trajectory and the probability density $p_B(\zeta_\tau^{xy}) = p(\widetilde{\zeta_\tau^{xy}})$ of the same couples of successive states ζ_τ^{xy} of the time-reversed conjugate of the original trajectory (Eq.5.8). For the sake of clarity, we use here in this appendix the full formalism rather than the compact one based on the functional form f_τ^{xy}.

The time-reversed density of a particular couple of successive states, $(x(t) = \gamma, y(t) = \delta)$ and $(x(t+\tau) = \mu, y(t+\tau) = \xi)$, is equivalent to the original density of the exchanged couple of states, $(x(t) = \mu, y(t) = \xi)$ and $(x(t+\tau) = \gamma, y(t+\tau) = \delta)$. Therefore the density $p(\widetilde{\zeta_\tau^{xy}}) = p(x(t) = \mu, y(t) = \xi, x(t+\tau) = \gamma, y(t+\tau) = \delta)$ is the transpose of the density $p(\zeta_\tau^{xy}) = p(x(t) = \gamma, y(t) = \delta, x(t+\tau) = \mu, y(t+\tau) = \xi)$.

The mapping irreversibility for the BLRM is then written as:

$$\begin{aligned} \Phi_\tau^{xy} = \langle \varphi_\tau^{xy} \rangle_{p(\zeta_\tau^{xy})} = \\ = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\gamma d\delta d\mu d\xi \; p\left(x(t)=\gamma, y(t)=\delta, x(t+\tau)=\mu, y(t+\tau)=\xi\right) * \\ * \ln\left(\frac{p(x(t)=\gamma, y(t)=\delta, x(t+\tau)=\mu, y(t+\tau)=\xi)}{p(x(t)=\mu, y(t)=\xi, x(t+\tau)=\gamma, y(t+\tau)=\delta)}\right). \end{aligned} \tag{5.23}$$

The BLRM is ergodic, therefore the densities $p(\zeta_\tau^{xy})$ and $p(\widetilde{\zeta_\tau^{xy}})$ can be empirically sampled looking at a single infinitely-long trajectory.

The causal structure of the BLRM (and of any signal-response model, see Fig.5.2) is such that the evolution of the signal is not influenced by the response, $p(x(t+\tau)|x(t), y(t)) = p(x(t+\tau)|x(t))$. Then we can write the joint probability densities $p(\zeta_\tau^{xy})$ of couples of successive states over a time interval τ of the original trajectory as:

$$\begin{aligned} p(\zeta_\tau^{xy}) \equiv p(x(t)=\gamma, y(t)=\delta, x(t+\tau)=\mu, y(t+\tau)=\xi) = \\ = p(x(t)=\gamma) \cdot p(y(t)=\delta|x(t)=\gamma) \cdot p(x(t+\tau)=\mu|x(t)=\gamma) * \\ * p(y(t+\tau)=\xi|x(t)=\gamma, y(t)=\delta, x(t+\tau)=\mu). \end{aligned} \tag{5.24}$$

We need to evaluate all these probabilities. Since we are dealing with linear models, these are all Gaussian distributed, and we will calculate only the expected value and the variance of the relevant variables involved.

The system is Markovian, $p(x(t+\tau)|x(t+t'),x(t)) = p(x(t+\tau)|x(t+t'))$ with $0 \leq t' \leq \tau$, and the Bayes rule assumes the form $p(x(t+t')|x(t),x(t+\tau)) = \frac{p(x(t+t')|x(t))p(x(t+\tau)|x(t+t'))}{p(x(t+\tau)|x(t))}$. Then we calculate the conditional expected value for the signal $x(t+\tau)$ given a condition for its past $x(t)$ and another condition for its future $x(t+\tau)$ as:

$$\langle x(t+t')|x(t),x(t+\tau)\rangle = x(t)e^{-\frac{t'}{t_{rel}}}\frac{1-e^{-\frac{2(\tau-t')}{t_{rel}}}}{1-e^{-\frac{2\tau}{t_{rel}}}} + x(t+\tau)e^{-\frac{\tau-t'}{t_{rel}}}\frac{1-e^{-\frac{2t'}{t_{rel}}}}{1-e^{-\frac{2\tau}{t_{rel}}}}. \quad (5.25)$$

Now we can calculate the full-conditional expectation of the response:

$$\langle y(t+\tau)|x(t),y(t),x(t+\tau)\rangle = y(t)e^{-\beta\tau} + \alpha\int_0^\tau dt' e^{-\beta(\tau-t')}\langle x(t+t')|x(t),x(t+\tau)\rangle = \quad y(t)e^{-\beta\tau} +$$
$$+\alpha\frac{e^{-\beta\tau}}{1-e^{-\frac{2\tau}{t_{rel}}}}\left(x(t)\left(\frac{e^{\tau(\beta-\frac{1}{t_{rel}})}-1}{\beta-\frac{1}{t_{rel}}} - \frac{e^{\tau(\beta-\frac{1}{t_{rel}})}-e^{-\frac{2\tau}{t_{rel}}}}{\beta+\frac{1}{t_{rel}}}\right) + x(t+\tau)\left(\frac{e^{\beta\tau}-e^{-\frac{\tau}{t_{rel}}}}{\beta+\frac{1}{t_{rel}}} - \frac{e^{\tau(\beta-\frac{2}{t_{rel}})}-e^{-\frac{\tau}{t_{rel}}}}{\beta-\frac{1}{t_{rel}}}\right)\right). \quad (5.26)$$

One can immediately check that the limits for small and large time intervals τ verify respectively $\lim_{\tau\to 0}\langle y(t+\tau)|x(t),y(t),x(t+\tau)\rangle = y(t)$ and $\lim_{\tau\to\infty}\langle y(t+\tau)|x(t),y(t),x(t+\tau)\rangle = x(t+\tau)\frac{\alpha t_{rel}}{\beta t_{rel}+1} = \langle y(t+\tau)|x(t+\tau)\rangle$.

The causal order for the evolution of the signal is such that $p(x(t+t'')|x(t),x(t+t'),x(t+\tau)) = p(x(t+t'')|x(t+t'),x(t+\tau))$ if $0 \leq t' \leq t'' \leq \tau$. Then we can calculate:

$$\langle x(t+t')x(t+t'')|x(t),x(t+\tau)\rangle_{t''\geq t'} =$$
$$= \int_{-\infty}^{\infty} dx(t+t')p(x(t+t')|x(t),x(t+\tau))x(t+t')\langle x(t+t'')|x(t+t'),x(t+\tau)\rangle =$$
$$= \langle x(t+t')|x(t),x(t+\tau)\rangle *$$
$$*\left(x(t+\tau)e^{-\frac{\tau-t''}{t_{rel}}}\frac{\sigma^2_{t''-t'}}{\sigma^2_{\tau-t'}} + e^{-\frac{t''-t'}{t_{rel}}}\frac{\sigma^2_{\tau-t''}}{\sigma^2_{\tau-t'}}\left(\frac{1}{x(t)\frac{e^{-\frac{t'}{t_{rel}}}}{\sigma^2_{t'}}+x(t+\tau)\frac{e^{-\frac{\tau-t'}{t_{rel}}}}{\sigma^2_{\tau-t'}}} + \frac{x(t)\frac{e^{-\frac{t'}{t_{rel}}}}{\sigma^2_{t'}}+x(t+\tau)\frac{e^{-\frac{\tau-t'}{t_{rel}}}}{\sigma^2_{\tau-t'}}}{\frac{1}{\sigma^2_{t'}}+\frac{e^{-\frac{2(\tau-t')}{t_{rel}}}}{\sigma^2_{\tau-t'}}}\right)\right). \quad (5.27)$$

Let us write the full-conditional expectation of the squared response as a function of the expectations we just calculated:

$$\langle y^2(t+\tau)|x(t),y(t),x(t+\tau)\rangle = y^2(t)e^{-2\beta\tau} + 2\alpha y(t)e^{-2\beta\tau}\int_0^\tau dt' e^{\beta t'}\langle x(t+t')|x(t),x(t+\tau)\rangle +$$
$$+\alpha^2 e^{-2\beta\tau}\int_0^\tau\int_0^\tau dt'dt'' e^{\beta(t'+t'')}\langle x(t+t')x(t+t'')|x(t),x(t+\tau)\rangle. \quad (5.28)$$

A relevant feature of linear response models is that the conditional variances do not depend on the particular values of the conditioning variables[Auc+17]. Here we consider the full-conditional variance $\sigma^2_{y(t+\tau)|x(t),y(t),x(t+\tau)} = \langle y^2(t+\tau)|x(t),y(t),x(t+\tau)\rangle -$

$\langle y(t+\tau)|x(t),y(t),x(t+\tau)\rangle^2$, and it will be independent of the conditions $x(t)$, $y(t)$, and $x(t+\tau)$. Then the remaining terms in $\sigma^2_{y(t+\tau)|x(t),y(t),x(t+\tau)}$ sum up to:

$$\begin{aligned}\sigma^2_{y(t+\tau)|x(t),y(t),x(t+\tau)} &= 2\frac{\alpha^2 e^{-2\beta\tau}}{\sigma^2_\tau}\int_0^\tau dt'' \sigma^2_{\tau-t''} e^{t''(\beta-\frac{1}{t_{rel}})}\int_0^{t''} dt' \sigma^2_{t'} e^{t'(\beta+\frac{1}{t_{rel}})} = \\ &= 2\alpha^2\sigma_x^2\frac{e^{-2\beta\tau}}{1-e^{-\frac{2\tau}{t_{rel}}}} * \Bigg(-\frac{2}{t_{rel}}\frac{\beta+\frac{1}{t_{rel}}-\frac{2}{t_{rel}}e^{\tau(\beta-\frac{1}{t_{rel}})}-(\beta-\frac{1}{t_{rel}})e^{-\frac{2\tau}{t_{rel}}}}{(\beta+\frac{1}{t_{rel}})^2(\beta-\frac{1}{t_{rel}})^2} + \\ &+\frac{\frac{1}{t_{rel}}e^{2\beta\tau}-\beta-\frac{1}{t_{rel}}+\beta e^{-\frac{2\tau}{t_{rel}}}}{2\beta(\beta+\frac{1}{t_{rel}})^2} - \frac{\frac{1}{t_{rel}}e^{2\tau(\beta-\frac{1}{t_{rel}})}-\beta+(\beta-\frac{1}{t_{rel}})e^{-\frac{2\tau}{t_{rel}}}}{2\beta(\beta-\frac{1}{t_{rel}})^2}\Bigg), \end{aligned} \tag{5.29}$$

where we used the fact that $\langle x(t+t')x(t+t'')|x(t),x(t+\tau)\rangle$ is symmetric in t' and t''. We recall that for functions with the symmetry $f(t',t'') = f(t'',t')$ it holds: $\int_0^\tau\int_0^\tau dt'dt'' f(t',t'') = 2\int_0^\tau dt'\int_0^{t'} dt'' f(t',t'')$.

The limits for small and large time intervals τ verify respectively $\lim_{\tau\to 0}\sigma^2_{y(t+\tau)|x(t),y(t),x(t+\tau)} = 0$ and $\lim_{\tau\to\infty}\sigma^2_{y(t+\tau)|x(t),y(t),x(t+\tau)} = \alpha^2\sigma_x^2\frac{t_{rel}}{\beta(\beta t_{rel}+1)^2} = \sigma^2_{y(t)|x(t)}$.

The factorization of the probability density $p(\zeta_\tau^{xy})$ into conditional densities (Eq.5.24) leads to a decomposition of the mapping irreversibility. Here we show that in the BLRM all of these terms are zero except for the two terms corresponding to the full-conditional density of the evolution of the response in the original trajectory and in the time-reversed conjugate.

For a stationary stochastic process like the BLRM it holds $p(x(t)=\gamma, y(t)=\delta) = p(x(t+\tau)=\gamma, y(t+\tau)=\delta)$, then these two terms cancel:

$$\begin{aligned}&\int_{-\infty}^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty d\gamma d\delta d\mu d\xi\ p(x(t)=\gamma,y(t)=\delta,x(t+\tau)=\mu,y(t+\tau)=\xi) * \\ &\qquad * \ln(p(x(t)=\gamma,y(t)=\delta)) = \\ &= \int_{-\infty}^\infty\int_{-\infty}^\infty d\gamma d\delta\ p(x(t)=\gamma,y(t)=\delta)\cdot\ln(p(x(t)=\gamma,y(t)=\delta)) = \\ &= \int_{-\infty}^\infty\int_{-\infty}^\infty d\gamma d\delta\ p(x(t+\tau)=\gamma,y(t+\tau)=\delta)\cdot\ln(p(x(t)=\gamma,y(t)=\delta)) = \\ &= \int_{-\infty}^\infty\int_{-\infty}^\infty d\mu d\xi\ p(x(t+\tau)=\mu,y(t+\tau)=\xi)\cdot\ln(p(x(t)=\mu,y(t)=\xi)) = \\ &= \int_{-\infty}^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty d\gamma d\delta d\mu d\xi\ p(x(t)=\gamma,y(t)=\delta,x(t+\tau)=\mu,y(t+\tau)=\xi) * \\ &\qquad * \ln(p(x(t)=\mu,y(t)=\xi)). \end{aligned} \tag{5.30}$$

The contribution from the signal in the mapping irreversibility is also zero since the Ornstein-Uhlenbeck process is reversible, $p(x(t)=\gamma,x(t+\tau)=\mu) = p(x(t)=\mu,x(t+\tau)=\gamma)$:

$$\int_{-\infty}^\infty\int_{-\infty}^\infty d\gamma d\mu\ p(x(t)=\gamma,x(t+\tau)=\mu)\ \ln\left(\frac{p(x(t+\tau)=\mu|x(t)=\gamma)}{p(x(t+\tau)=\gamma|x(t)=\mu)}\right) = 0. \tag{5.31}$$

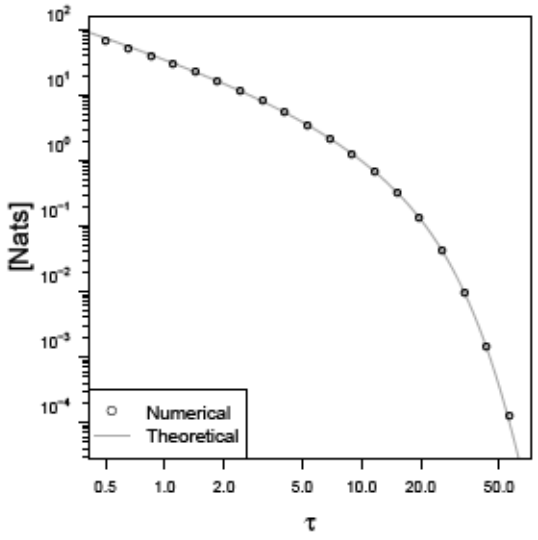


Fig. 5.8: Numerical verification of the analytical solution for the entropy production Φ_τ^{xy} with observational time τ in the BLRM. The parameters are $\beta = 0.2$ and $t_{rel} = 10$. The slight down-deviation for small τ is due to the finite box length in the discretized space, while the up-deviation for $\tau \to \infty$ is due to the finite number of samples.

The mapping irreversibility is therefore:

$$\begin{aligned}\Phi_\tau^{xy} = \langle \varphi_\tau^{xy} \rangle_{p(\zeta_\tau^{xy})} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\gamma d\delta d\mu d\xi \; p(x(t)=\gamma, y(t)=\delta, x(t+\tau)=\mu, y(t+\tau)=\xi) * \\ * \ln\left(\frac{p(y(t+\tau)=\xi | x(t)=\gamma, y(t)=\delta, x(t+\tau)=\mu)}{p(y(t+\tau)=\delta | x(t)=\mu, y(t)=\xi, x(t+\tau)=\gamma)}\right) = \\ = \frac{1}{2\sigma^2_{y(t+\tau)|x(t),y(t),x(t+\tau)}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\gamma d\delta d\mu d\xi \; p(x(t)=\gamma, y(t)=\delta, x(t+\tau)=\mu, y(t+\tau)=\xi) * \\ *(\delta - \langle y(t+\tau) | x(t)=\mu, y(t)=\xi, x(t+\tau)=\gamma \rangle)^2 \; - \tfrac{1}{2}, \end{aligned} \tag{5.32}$$

where in the last passage we exploited the fact that all the probability densities are Gaussian distributed. Solving the integrals we get the mapping irreversibility for the BLRM as a function of the time interval τ:

$$\Phi_\tau^{xy} = \frac{1}{2}(e^{-2\beta\tau} - 1) + \frac{2\alpha^2\sigma_x^2 t_{rel}^2}{\sigma^2_{y(t+\tau)|x(t),y(t),x(t+\tau)}} \frac{\left(e^{-\beta\tau} - e^{-\frac{\tau}{t_{rel}}}\right)^2 \left(e^{-2\beta\tau} + 1 - 2e^{-\tau(\beta+\frac{1}{t_{rel}})}\right)}{(\beta^2 t_{rel}^2 - 1)^2 \left(1 - e^{-\frac{2\tau}{t_{rel}}}\right)}. \tag{5.33}$$

$\sigma^2_{y(t+\tau)|x(t),y(t),x(t+\tau)}$ is proportional to $\alpha^2\sigma_x^2$, therefore the mapping irreversibility Φ_τ^{xy} is a function of just the two parameters t_{rel} and β (and of the observational time τ). Since we are allowed to change the units for time, the shape of Φ_τ^{xy} result to be dependent on a single parameter being the product βt_{rel}.

5.4.2 Appendix B: Backward transfer entropy in the BLRM

In the BLRM, where all densities are Gaussian distributed, the backward transfer entropy is equivalent to the time-reversed Granger causality[Bar+09]:

$$T_{y\to x}(-\tau) = I(x(t), y(t+\tau)|x(t+\tau)) = \ln\left(\frac{\sigma_{y|x}}{\sigma_{y(t+\tau)|x(t),x(t+\tau)}}\right). \tag{5.34}$$

We have to calculate the conditional variance $\sigma_{y(t+\tau)|x(t),x(t+\tau)}$. Let us recall the relation for the value of the response as a function of the whole past history of the signal trajectory:

$$y(t+\tau) = \alpha e^{-\beta(t+\tau)}\left(\int_{-\infty}^{t} dt' x(t') e^{\beta t'} + \int_{t}^{t+\tau} dt' x(t') e^{\beta t'}\right). \tag{5.35}$$

Then we write the conditional squared response as

$$\begin{aligned}
&\langle y^2(t+\tau)|x(t), x(t+\tau)\rangle = \\
&= 2\alpha^2 e^{-2\beta(t+\tau)}\Big(e^{2\beta t}\int_0^{\tau}\int_0^{t''} dt''dt' e^{\beta(t'+t'')}\,\langle x(t+t')x(t+t'')|x(t), x(t+\tau)\rangle_{t''\geq t'} + \\
&\quad + \int_{-\infty}^{t}\int_{-\infty}^{t''} dt''dt'\,\langle x^2(t'')|x(t)\rangle\, e^{-\frac{t''-t'}{t_{rel}}+\beta(t'+t'')} + \\
&\quad + \int_{-\infty}^{t}\int_{t}^{t+\tau} dt'dt''\,\langle x(t')|x(t)\rangle\,\langle x(t'')|x(t), x(t+\tau)\rangle\, e^{\beta(t'+t'')}\Big).
\end{aligned} \tag{5.36}$$

Since $\sigma^2_{y(t+\tau)|x(t),x(t+\tau)}$ is expected to be independent of $x(t)$ and $x(t+\tau)$, then the remaining terms sum up to:

$$\sigma^2_{y(t+\tau)|x(t),x(t+\tau)} = \sigma^2_{y(t+\tau)|x(t),y(t),x(t+\tau)} + \sigma^2_{y|x} e^{-2\beta\tau}, \tag{5.37}$$

where $\sigma^2_{y|x} = \frac{\sigma_x^2\alpha^2}{\beta t_{rel}(\beta+\frac{1}{t_{rel}})^2}$ was already calculated in chapter 4 (and in [Auc+17]). Then the backward transfer entropy is:

$$T_{y\to x}(-\tau) = I\left(x(t), y(t)|x(t+\tau)\right) = -\frac{1}{2}\ln\left(\frac{\sigma^2_{y(t+\tau)|x(t),y(t),x(t+\tau)}}{\sigma^2_{y|x}} + e^{-2\beta\tau}\right). \tag{5.38}$$

5.4.3 Appendix C: The causal influence rate converges to the Horowitz-Esposito information flow in the BLRM

We introduced the Horowitz-Esposito information flow[HS14; HE14] in Eq.5.4. In our stationary processes framework, the two components of the information flow are related by $I_{x\to y} = -I_{y\to x}$, so that the information flow is unidirectional and necessarily asymmetric when present. The y variable in the BLRM is measuring the

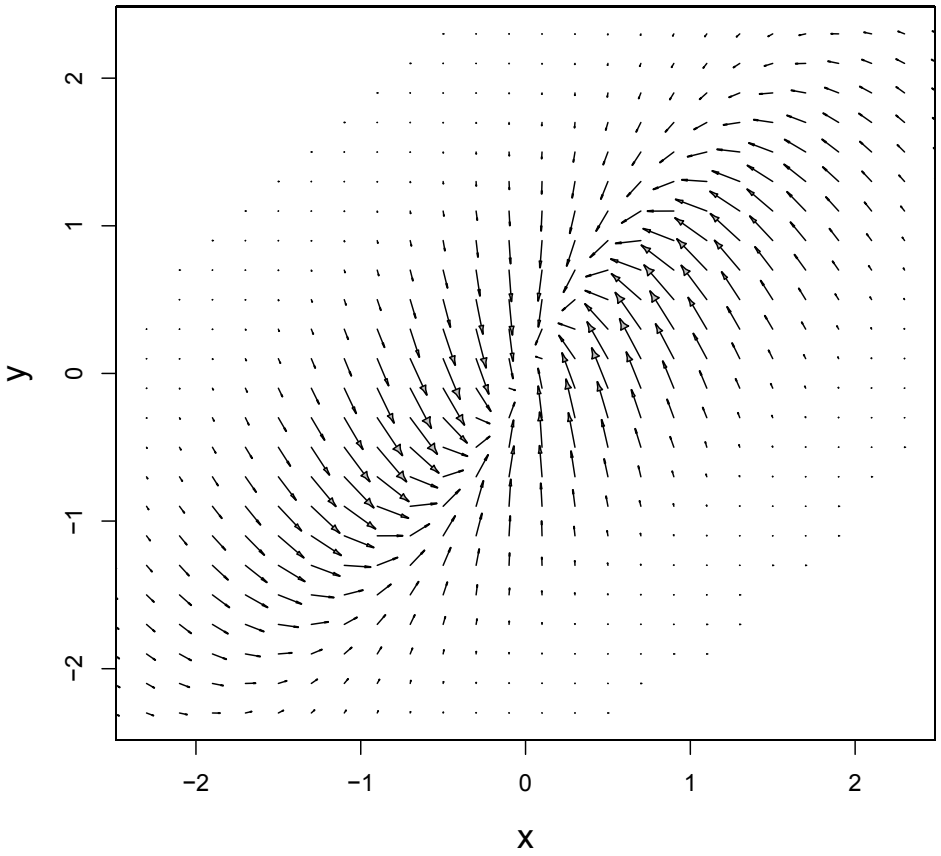


Fig. 5.9: Probability currents $\vec{J}$ in the BLRM, estimated with $\tau = 0.1$. The parameters are $\beta = 0.2$ and $t_{rel} = 10$. The space coordinates are in units of the standard deviation.

x variable, therefore the information is flowing in the $x \to y$ direction, and we wish to calculate the parameter dependence of the positive $I^{x\to y}$:

$$I^{x\to y} = \int\int dxdy\ J_y(x,y,t)\frac{\frac{\partial}{\partial y}p(x|y)}{p(x|y)}. \tag{5.39}$$

We see that the information flow is a function of probability currents. We plot the current $\vec{J}$ for the BLRM in Fig.5.9 The probability flow $J_y(x,y,t)$ in the y direction for the BLRM is given by $J_y = (\alpha x - \beta y)p(x,y)$. Then we calculate:

$$\begin{aligned} I^{x\to y} &= \int\int dxdy\ p(y)(\alpha x - \beta y)\tfrac{\partial p(x|y)}{\partial y} = \beta - \int\int dxdy\ p(x|y)(\alpha x - \beta y)\tfrac{\partial p(y)}{\partial y} = \\ &= \beta + \int\int dxdy(\alpha x - \beta y)p(x,y)\tfrac{y}{\sigma_y^2} = \tfrac{\alpha}{\sigma_y^2}\left\langle xy\right\rangle, \end{aligned} \tag{5.40}$$

where in the second passage we used partial integration and $p(y)(\alpha x - \beta y)$ is exponentially vanishing for $y \to \pm\infty$ because $p(y)$ is a Gaussian, $p(y) = N(0,\sigma_y^2)$.

In the last passage we identified $\int\int dxdy\ p(x,y)xy = \langle xy \rangle$. Since the BLRM is a stationary process the time derivatives of expectations vanish, then $0 = \frac{d\langle x(t)y(t)\rangle}{dt} = \left\langle \frac{dx}{dt} y \right\rangle + \left\langle x \frac{dy}{dt} \right\rangle = -\langle xy \rangle (\beta + \frac{1}{t_{rel}}) + \alpha\sigma_x^2$, and we find $\langle xy \rangle = \frac{\alpha\sigma_x^2}{\beta + \frac{1}{t_{rel}}}$. Then using the BLRM expression (see Chapter 4 or [Auc+17]) for the variance of the response $\sigma_y^2 = \frac{\alpha^2 t_{rel}}{\beta(\beta t_{rel}+1)}\sigma_x^2$, we obtain:

$$I^{x\to y} = \beta. \tag{5.41}$$

The Horowitz-Esposito information flow in the BLRM is equal to the inverse of the deterministic response time to perturbations $\frac{1}{\beta}$. Interestingly, this is independent of the time scale of fluctuations t_{rel}. Let us consider a fixed β, then if t_{rel} is small we have very fast fluctuations and the response is not able to follow the signal with accuracy and the mutual information $I(x,y) = \frac{1}{2}\ln(1+\beta t_{rel})$ is small. Nevertheless the information flow $I^{x\to y}$ does not decrease because the dynamics of the y variable is driven by the x position for every possible situation (x,y) even if not strongly correlated.

Importantly, in the small observational time limit our definition of causal influence[Auc+17] converges in rate to the Horowitz-Esposito information flow:

$$\lim_{\tau\to 0} \frac{C_{x\to y}(\tau)}{\tau} = I^{x\to y}. \tag{5.42}$$

5.4.4 Appendix D: Numerical convergence of the mapping irreversibility to the entropy production in the feedback cooling model

The feedback cooling model[RH16; HS14] describes a Brownian particle with velocity x and viscous damping $\gamma > 0$, that is under the feedback control of the measurement device y. The variable y is a a low-pass filter of noisy measurements on x. The SDE system describing the process is written:

$$\begin{cases} dx = -(\gamma x + ky)\ dt + \sqrt{D_x}\ dW_x \\ dy = (x - y)\ dt + \sqrt{D_y}\ dW_y \end{cases} \tag{5.43}$$

$k > 0$ is the feedback coefficient, while dW_x and dW_y are uncorrelated Brownian noise sources. The mapping irreversibility (5.8) converges in the limit of small observational time $\tau \to 0$ to the physical entropy production (5.5) if the conditional probability $p(x_{t+\tau}, y_{t+\tau}|x_t, y_t) = p(x_{t+\tau}|x_t, y_t) \cdot p(y_{t+\tau}|x_t, y_t, x_{t+\tau})$ converges almost surely to the bipartite form $p(x_{t+\tau}|x_t, y_t) \cdot p(y_{t+\tau}|x_t, y_t)$. Importantly, the convergence has to be faster than τ so that in the limit of continuous trajectories we can almost surely neglect the term $\lim_{\tau\to 0} \frac{1}{\tau} \ln \frac{p(y_{t+\tau}|x_t, y_t, x_{t+\tau})}{p(y_{t+\tau}|x_t, y_t)} = 0$.

The knowledge of $x_{t+\tau}$ acts only on the estimate of $f_y(x_{t+t'}, y_{t+t'})$ (with $0 \leq t' \leq \tau$) because the diffusion coefficients are constant. Since the system (5.43) is linear, the Kullback-Leibler divergence can be expressed in terms of conditional expectations:

$$\left\langle \ln \frac{p(y_{t+\tau}|x_t,y_t,x_{t+\tau})}{p(y_{t+\tau}|x_t,y_t)} \right\rangle_{p(\zeta_\tau^{xy})} =$$
$$= -\ln \frac{\sigma_{y_{t+\tau}|x_t,y_t,x_{t+\tau}}}{\sigma_{y_{t+\tau}|x_t,y_t}} + \frac{\sigma^2_{y_{t+\tau}|x_t,y_t,x_{t+\tau}} + \left\langle (\langle y_{t+\tau}|x_t,y_t,x_{t+\tau}\rangle - \langle y_{t+\tau}|x_t,y_t\rangle)^2 \right\rangle_{p(\zeta_\tau^{xy})}}{2\sigma^2_{y_{t+\tau}|x_t,y_t}} - \frac{1}{2}. \qquad (5.44)$$

The conditional variance $\sigma^2_{y_{t+\tau}|x_t,y_t,x_{t+\tau}}$ is of order $\langle W_y^2(\tau)\rangle \sim \tau$ and differs from $\sigma^2_{y_{t+\tau}|x_t,y_t}$ only with a term of order $(\partial_x f_y(x,y))^2 \langle W_x^2(\tau)\rangle \tau^2 \sim \tau^3$ so that $\ln \frac{\sigma_{y_{t+\tau}|x_t,y_t,x_{t+\tau}}}{\sigma_{y_{t+\tau}|x_t,y_t}} \sim \tau^2$ for $\tau \to 0$.

While the conditional variances are constant, the conditional expectation $\langle y_{t+\tau}|x_t, y_t, x_{t+\tau}\rangle$ depend linearly on $x_{t-\tau}$ (and on x_t, y_t), therefore it is sufficient to look at the conditional correlation $C(x_{t+\tau}, y_{t+\tau}|x_t, y_t) = \frac{\langle x_{t+\tau}y_{t+\tau}|x_t,y_t\rangle - \langle x_{t+\tau}|x_t,y_t\rangle\langle y_{t+\tau}|x_t,y_t\rangle}{\sigma_{x_{t+\tau}|x_t,y_t}\sigma_{y_{t+\tau}|x_t,y_t}}$, given that $\left\langle \frac{(\langle y_{t+\tau}|x_t,y_t,x_{t+\tau}\rangle - \langle y_{t+\tau}|x_t,y_t\rangle)^2}{\sigma_{y_{t+\tau}|x_t,y_t}} \right\rangle_{p(x_{t+\tau}|x_t,y_t)} = C^2(x_{t+\tau}, y_{t+\tau}|x_t, y_t)$. By numerical simulation we checked that $\frac{1}{\tau}C^2(x_{t+\tau}, y_{t+\tau}|x_t, y_t) \to 0$ in the limit $\tau \to 0$ for the feedback cooling model (5.43). We also checked that with $D_y = 0$ there is no convergence.

For the case of nonconstant diffusion coefficients (moltiplicative noise) the argument on the conditional variances does not hold, and we are not sure of the convergence.

5.4.5 Appendix E: Numerical estimation of the entropy production in the bivariate Gaussian approximation

We calculate numerically the mapping irreversibility Φ_τ^{xy} as an average of the spatial density of entropy irreversibility $\psi(x_t, y_t)$, $\Phi_\tau^{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx_t dy_t \psi(x_t, y_t)$. In the computer algorithm the (x, y) space is dicretized in boxes (i, j), and for each box we estimate the conditional correlation $C_{xy|i,j}$ of future values $(x_{t+\tau}, y_{t+\tau})$, the conditional correlation $C_{xy|i,j}$ of past values $(x_{t-\tau}, y_{t-\tau})$, the expected values for both variables in future ($\langle x|i,j\rangle$, $\langle y|i,j\rangle$) and past states ($\langle \widetilde{x}|i,j\rangle$, $\langle \widetilde{y}|i,j\rangle$), and the standard deviations on those $\sigma_{x|i,j}$, $\sigma_{x|i,j}$, $\sigma_{y|i,j}$, $\sigma_{y|i,j}$. The spacial density evaluated

in the box (i,j) is then calculated as the bidimensional Kullback-Leibler divergence in the Gaussian approximation[Duc07]:

$$\psi(i,j) = P(i,j)\ \frac{1}{2}\left[\ln\left(\frac{\sigma^2_{x|i,j}\sigma^2_{y|i,j}(1-C^2_{xy|i,j})}{\sigma^2_{x|i,j}\sigma^2_{y|i,j}(1-C^2_{xy|i,j})}\right) - 2 + \frac{\frac{\sigma^2_{x|i,j}}{\sigma^2_{x|i,j}} + \frac{\sigma^2_{y|i,j}}{\sigma^2_{y|i,j}} - 2\frac{\sigma_{x|i,j}\sigma_{y|i,j}}{\sigma_{x|i,j}\sigma_{y|i,j}}C_{xy|i,j}C_{xy|i,j}}{1-C^2_{xy|i,j}} + \right.$$
$$\left. + \frac{\sigma^2_{y|i,j}(\langle\widetilde{x}|i,j\rangle - \langle x|i,j\rangle)^2 + \sigma^2_{x|i,j}(\langle\widetilde{y}|i,j\rangle - \langle y|i,j\rangle)^2 - 2C_{xy|i,j}\sigma_{x|i,j}\sigma_{y|i,j}(\langle\widetilde{x}|i,j\rangle - \langle x|i,j\rangle)(\langle\widetilde{y}|i,j\rangle - \langle y|i,j\rangle)}{\sigma^2_{x|i,j}\sigma^2_{y|i,j}(1-C^2_{xy|i,j})}\right]. \quad (5.45)$$

The effect of the finite width of the discretization is attenuated by estimating all the quantities taking into account the starting point (x_t, y_t) within the box (i,j), subtracting the difference to the mean values for each box. For example, when we sample for the estimate of the conditional average $\langle x_{t+\tau}|i\rangle$ we would collect samples $x_{t+\tau} - (x_t - \langle x_t|i\rangle)$.

5.4.6 Appendix F: Lower bound and statistics

We claimed that, when only a small finite number of samples is available, like having a single short time series, the backward transfer entropy $T_{y\to x}(\tau)$ can be estimated with more precision than the mapping irreversibility Φ^{xy}_τ. This is because Φ^{xy}_τ is calculated from the statistics of 4 variables $(x_t, y_t, x_{t+\tau}, y_{t+\tau})$, while $T_{y\to x}(\tau)$ is calculated from the statistics of 3 variables $(x_t, x_{t+\tau}, y_{t+\tau})$.

Here is the numerical evidence for this in the BLRM for which we have the actual values in analytical form. We calculate Φ^{xy}_τ partitioning the (x_t, y_t) space in N^2_{box} boxes, and assuming that in each box the conditional distributions $p(x_{t+\tau}, y_{t+\tau}|x_t, y_t)$ and $p(x_{t-\tau}, y_{t-\tau}|x_t, y_t)$ are both bivariate Gaussians. Then we use the formula for the KL divergence between 2D Gaussians (see Appendix E) and sum up the irreversibility contributions of the relevant boxes (those with at least 3 samples in it). Then we estimate $T_{y\to x}(\tau)$ partitioning the $x_{t+\tau}$ space in (only) N_{box} boxes, evaluating the conditional correlation $C(x_t, y_{t+\tau}|x_{t+\tau})$ for each $x_{t+\tau}$ box, and then computing the weighted sum of the terms $-\frac{1}{2}\ln(1 - C^2(x_t, y_{t+\tau}|x_{t+\tau}))$.

Choosing a dynamics with $t_{rel} = 10$, we take 50 replicas of a time series of length $T = 100 \cdot t_{rel} = 1000$, sampled with interval $\Delta t = \frac{t_{rel}}{100} = 0.1$. We estimated the quantities in each of the replicas separately and plotted those in Fig.5.10 together with the analytical solutions. It is clearly seen how better the backward transfer entropy is estimated compared to the mapping irreversibility. There we choose $N_{box} = 15$, but we verified that the plot is qualitatively the same also for $N_{box} = 5$ and $N_{box} = 50$.

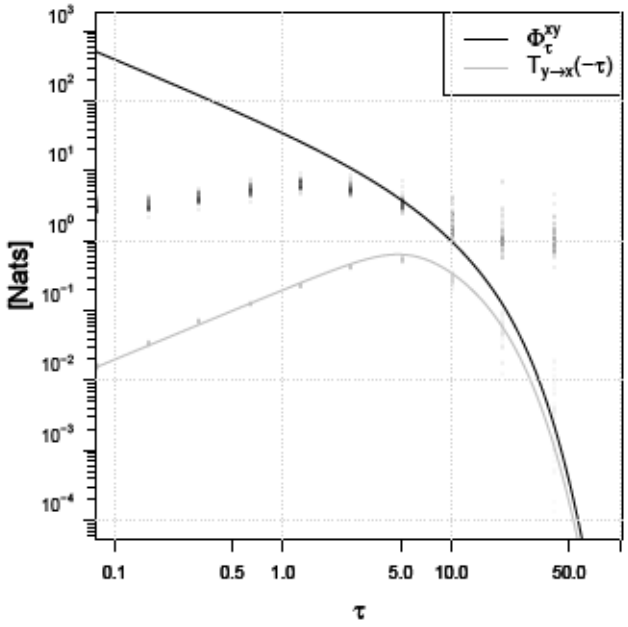


Fig. 5.10: 50 replicas of a numerical estimation experiment (points) from short time series. The backward transfer entropy $T_{y\to x}(-\tau)$ converges faster to the analytical solution (line), compared to the mapping irreversibility Φ_τ^{xy}.

6 Quantifying the effect of photic perturbations on circadian rhythms.

This Chapter introduces preliminary results of an application of the time series irreversibility framework developed in Chapter 5 to circadian theoretical biology. These ideas are the fruit of a collaboration with Prof. Hanspeter Herzel and Dr. Patrick J Pett at the Institute of Theoretical Biology in Berlin, and have been formalized in a manuscript available in the Arxiv[Auc+19a], and currently under review for publication.

6.1 Circadian clock biology

The circadian clock is a set of endogenous molecular mechanisms that allow adaptation to daily rhythms[Fuh+15]. The field of Chronobiology[Dun+04] has developed as an interdisciplinary science, where the mechanistic interpretation of experimental results was motivated by the physicists' long standing interest in oscillations.

The central site of circadian rhythms is the suprachiasmatic nucleus (SCN) located in the brain hypothalamus , and it synchronizes secondary clocks throughout the body. Oscillations are produced at the single-cell level by a network of gene transcriptional regulation whose main components, the circadian genes, are well identified in mammals[Bor+09; Leh+15]. Dysfunctions of the circadian clock are involved in diseases like sleep disorder[Van+06], but also cancer[SSC09] and diabetes[Mar+10].

The circadian clock is coupled to the external day/night cycle through the input signalling pathways that process the timing cues or zeitgebers, the main one being light. Light signals are perceived in the retina and transmitted to the SCN where the interaction of glutamate with NMDA receptors promote calcium influx. The circadian clock is entrained by light, but is self-sustained, meaning that circadian rhythm persist in the absence of external inputs, like in constant dark experiments.

We are interested in the mammalian core-clock network, that is the gene regulatory network that generates robust oscillations in single cells of our body. Mathemat-

ical models were developed to formalize and interpret biological knowledge and quantitative data[Kor+14; Rel+11; LG03; FP03]. These are differential equations models (ODE, time-delayed DDE, or stochastic SDE), and the general consensus feature is the multiple feedback structure[BW+04]. The heterodimer complex $CLOCK/BMAL1$ binds to the promoter region E-Box sequences of target genes $Rev\text{-}Erb$, Per, Ror, and Cry. PER/CRY complexes translocate into the nucleus and inhibit $CLOCK/BMAL1$ mediated transcription creating the main oscillating feedback loop. A second feedback loop with $Rev\text{-}Erb$ and Ror is interconnected to the PER/CRY loop, and it ensures 24h robust rhythms[Rel+11].

A large portion of genes are under circadian control, 50% was reported in mouse tissues[Zha+14], therefore most important processes are regulated by the circadian clock including metabolism, memory consolidation, blood pressure and body temperature[Maz+12; Fuh+15].

The mouse is the preferred experimental model because of its similarity to the human circadian structure. Circadian phenotype modifications are studied under perturbations of light/dark cycles or temperature[Abr+18]. The transcriptional interaction between genes is studied with classical bioinformatics analysis of ChIP-sequencing data[Koi+12].

Our main interest is the circadian entrainment, that is the set of environmental time cues that regulate the mammalian circadian clock, and in particular we focus on the photic entrainment. The temporal adaptation given by the circadian entrainment allows an accurate anticipation in the organism of the natural periodic changes. Light pulses are particularly effective in influencing the phase of the circadian clock during the boundary hours between the subjective day and night[GR10].

The fundamental description of the circadian system's sensitivity to light fluctuations is the phase response curve (PRC)[Gra+09; Joh99]. Let us consider nonlinear dynamical systems characterized by stable limit cycles, hereafter called self-sustained oscillations. Phase-response curves describe the phase response of limit cycle oscillations to single pulse-like perturbations as a function of the phase within the cycle at which the perturbation is applied. Similarly, the phase transition curve is defined as the mapping between the old phase and new phase after the pulse, $\phi_{n+1} = f(\phi_n)$.

We will study the response to photic entrainment in a dynamic out-of-equilibrium setting, meaning for a continuous perturbation. We will use the time series stochastic thermodynamics framework developed in chapter 5, and quantify the lag-dependent response with an information-theoretic measure called mutual mapping irreversibility (a Markovian approximation to the previously defined mutual entropy production[DE14]).

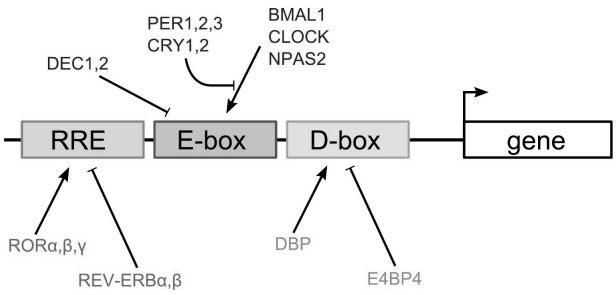


Fig. 6.1: Schematic representation of transcriptional regulation on promoter elements. The plot is taken from [Kor+14].

6.2 The circadian clock model

Our aim is the characterization of the influence of photic perturbations on circadian oscillations. Then we need to select one of the many mathematical models present in the literature, and discuss the appropriate way to add the influence of photic perturbations. Our physics background makes us always decide for the minimal model, with the condition that it should describe most of all the experimental consensus *quantitative* information. Further knowledge on additional components from bioinformatic studies should be neglected in differential equation models if their interactions or regulations are poorly characterized. In particular, ODE models become easily chaotic in higher dimensions, while in practical applications (biology, engineering, physics) the use of mathematics is done in order to formalize quantitative features and make predictions obtaining closed form expressions, or developing nonlinear algorithms with ensured converge. Here chaotic has a precise meaning, that is the absence of a periodic, diverging or converging solutions[Vul10]. Chaotic systems, while being very interesting from a mathematical standpoint, and probably being the appropriate description of any real physical dynamics, are to be avoided in practical applications. The uncertainty on the data should be considered as a limitation, and not as a source of deterministic chaos in the assumed dynamics, this being a form of qualitative over-fitting. Uncertainty should instead be modeled with stochastic processes. In particular, it was shown to be impossible to differentiate chaos from stochasticity in real data[Cen+00]. A differential equation model should be constructed only when reasonable amount of quantitative information is available, otherwise the simpler Boolean models are to be preferred[Thi+17; RK17]. This is necessarily the case when a large number of components are present, and one wishes to represent all the available biological experimental non quantitative (only logic ON/OFF) knowledge.

We will consider the minimal model defined in [Kor+14], that is composed of 5 delay-differential equations. That is a data-driven model based on expression profiles

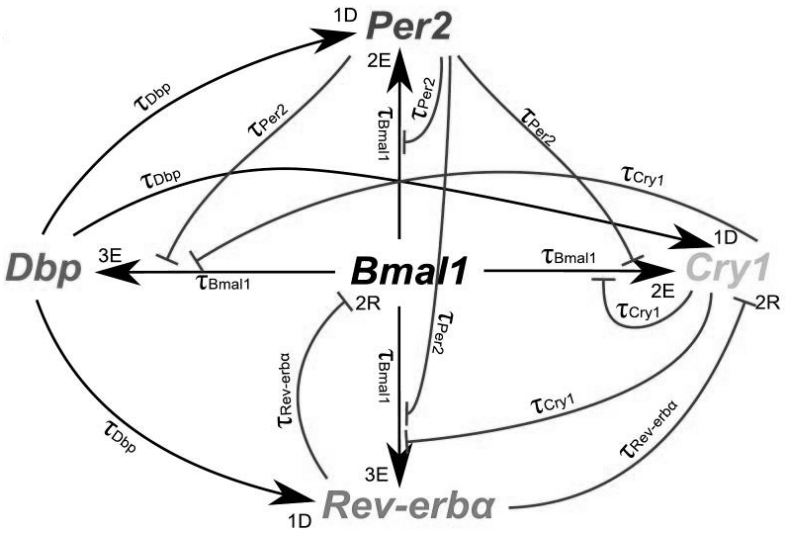


Fig. 6.2: Circadian clock model structure. The graph is directed because interactions are asymmetric. Such directed influences are exerted with explicit time delays representing the time lag between peaks of mRNA and proteins. Normal arrows indicate activation, while T-arrows indicate inhibitions. The plot is taken from [Kor+14].

and known transcriptional regulatory sites, these being represented in Fig.6.1. In particular, quantitative time-resolved PCR data are available under different conditions. Differences between light-dark and constant dark experiments were relatively small. The model is able to reproduce the correct phase-separation of transcriptional peaks within the cycle, and its variability among different tissues. Regulation is modeled in the multiplicative form, that was motivated by considerations of energy and configuration combinatorics in the equilibrium statistical mechanics framework[Bin+05]. Multiplicative regulation was already shown to exhibit 12h harmonics in circadian oscillations[WH13]. Exponents in the multiplicative terms represent the presence of multiple regulatory sites in the promoter region of the corresponding gene. The number of components in the model is reduced to just five, due to the merging of poorly characterized intermediate steps into delays in the asymmetric interactions between variables. Then events like phosphorylations, complex formation, and nuclear localization are not considered in detail but only as a single fixed time interval required for the interaction with a specific gene. The complex regulatory network is represented in Fig.6.2. We will report the DDE model hereafter along with the parameters, and it can also be found in [Kor+14]. In particular, we use the "consensus parameters" averaged over different tissues. Parameter values from different tissues are anyway on the same scale, and the main characteristics of oscillations like the order of peaks are preserved in all mammalian tissues. Let us just mention that small variations in the core-clock oscillations can lead to large variations in the output clock-dependent genes[Kor+14].

The deterministic circadian core clock model equations read:

$$\frac{dBmal1}{dt} = \left(1 + \frac{RevErb_{t-\tau_{RevErb}}}{ar1}\right)^{-2} - d_{Bmal1} Bmal1_t. \tag{6.1}$$

$$\frac{dRevErb}{dt} = \left(\frac{1+b_{RevErb} Bmal1_{t-\tau_{Bmal1}} \frac{1}{ba2}}{1+Bmal1_{t-\tau_{Bmal1}} \frac{1}{ba2}}\right)^3 \left(1 + Per2_{t-\tau_{Per2}} \frac{1}{cr2}\right)^{-3} \left(1 + Cry1_{t-\tau_{Cry1}} \frac{1}{gr2}\right)^{-3} * \\ * \frac{1+f_{RevErb} Dbp_{t-\tau_{Dbp}} \frac{1}{fa2}}{1+Dbp_{t-\tau_{Dbp}} \frac{1}{fa2}} - d_{RevErb} RevErb_t. \tag{6.2}$$

$$\frac{dPer2}{dt} = \left(\frac{1+b_{Per2} Bmal1_{t-\tau_{Bmal1}} \frac{1}{ba3}}{1+Bmal1_{t-\tau_{Bmal1}} \frac{1}{ba3}}\right)^2 \left(1 + Per2_{t-\tau_{Per2}} \frac{1}{cr3}\right)^{-2} \left(1 + Cry1_{t-\tau_{Cry1}} \frac{1}{gr3}\right)^{-2} * \\ * \left(\frac{1+f_Per2 Dbp_{t-\tau_{Dbp}} \frac{1}{fa3}}{1+Dbp_{t-\tau_{Dbp}} \frac{1}{fa3}}\right) - d_{Per2} Per2_t. \tag{6.3}$$

$$\frac{dCry1}{dt} = \left(\frac{1+b_{Cry1} Bmal1_{t-\tau_{Bmal1}} \frac{1}{ba4}}{1+Bmal1_{t-\tau_{Bmal1}} \frac{1}{ba4}}\right)^2 \left(1 + Per2_{t-\tau_{Per2}} \frac{1}{cr4}\right)^{-2} \left(1 + Cry1_{t-\tau_{Cry1}} \frac{1}{gr4}\right)^{-2} * \\ * \left(1 + RevErb_{t-\tau_{RevErb}} \frac{1}{ar4}\right)^{-2} \left(\frac{1+f_Cry1 Dbp_{t-\tau_{Dbp}} \frac{1}{fa4}}{1+Dbp_{t-\tau_{Dbp}} \frac{1}{fa4}}\right) - d_{Cry1} Cry1_t. \tag{6.4}$$

$$\frac{dDbp}{dt} = \left(\frac{1+b_{Dbp} Bmal1_{t-\tau_{Bmal1}} \frac{1}{ba5}}{1+Bmal1_{t-\tau_{Bmal1}} \frac{1}{ba5}}\right)^3 * \\ * \left(1 + Per2_{t-\tau_{Per2}} \frac{1}{cr5}\right)^{-3} \left(1 + Cry1_{t-\tau_{Cry1}} \frac{1}{gr5}\right)^{-3} - d_{Dbp} Dbp_t. \tag{6.5}$$

The consensus parameters are given by: $\tau_{Bmal1} = 4.76$; $\tau_{RevErb} = 1.79$; $\tau_{Per2} = 3.82$; $\tau_{Cry1} = 3.13$; $\tau_{Dbp} = 2.08$; $d_{Bmal1} = 0.46$; $d_{RevErb} = 0.67$; $d_{Per2} = 0.51$; $d_{Cry1} = 0.2$; $d_{Dbp} = 0.56$; $ar1 = 4.05$; $ar4 = 1.1$; $cr2 = 1.83$; $cr3 = 33.5$; $cr4 = 6.63$; $cr5 = 0.99$; $gr2 = 80.2$; $gr3 = 0.37$; $gr4 = 0.51$; $gr5 = 1.02$; $b_{RevErb} = 3.26$; $ba2 = 0.51$; $b_{Per2} = 3.69$; $ba3 = 14.78$; $b_{Cry1} = 1.35$; $ba4 = 1.06$; $b_{Dbp} = 12.87$; $ba5 = 0.01$; $fa2 = 0.19$; $f_{RevErb} = 1.23$; $fa3 = 0.58$; $f_{Per2} = 11.69$; $fa4 = 1.61$; $f_{Cry1} = 32.2$.

As you can see, the circadian core clock model is highly nonlinear and strongly characterized by feedback. Nevertheless, we can keep the signal-response structure if we study the response of such a model to an external perturbation, like a continuous fluctuating light entrainment. The results will still reflect the feedback properties of the model, and in particular its oscillatory behavior. Interestingly the response to light fluctuations, that we will quantify with a mutual irreversibility measure as discussed in the next section, will be strongly characterized by 12h harmonics.

6.3 Time series information thermodynamics of the perturbed circadian oscillations

Our starting point is the minimal model of [Kor+14] (Eq.6.1-6.5), that consists of delay differential equations for the five coarse-grained variables $\vec{y} \equiv (Bmal1, Per2, Cry1, Rev\text{-}erb\alpha, Dbp)$, each one representing more than just one gene transcript. These variables measure concentrations so they are taken to be positive, $y_i > 0 \ \forall i$. The dynamics produces deterministic self-sustained oscillations with a limit cycle, whose amplitudes and phase-differences between genes can be tuned to reproduce heterogeneity of different tissues. The dynamics is composed of degradation terms which are simply linear in the concentrations, and of production terms which are modeled as products of activation and repression functions of Michaelis-Menten type. While the model structure was based on biological knowledge, its parameters were optimized on experimental time-resolved quantitative data on mammalian tissues. Each variable y_i regulates the dynamics of other variables (and of its own) with a different time delay τ_i. Among the many interactions in the model, a network motif was identified as the main driving force of self-sustained oscillations[Pet+16], this being the repressilator loop of the three subsequent inhibitions $Per2 \dashv Rev\text{-}erb\alpha \dashv Cry1 \dashv Per2$.

Photic perturbations on the mammalian circadian system are perceived in the core suprachiasmatic nucleus through induction of $Per2$ genes, this mechanism being most sensitive during the night[GR10; Yan09; RW01]. Here we develop a framework based on time series irreversibility aspects for quantifying the influence of photic perturbations on circadian rhythms. Unlike phase-response curves[Gra+09], that evaluate the resulting phase-lag after the dynamics has relaxed to the limit cycle, we explicitly take into account dynamical aspects of the response to continuous perturbations keeping the system continuously out of equilibrium.

Obviously the light influence is an asymmetric interaction, meaning that the perturbation dynamics is not affected by the circadian dynamics, and all the feedbacks are endogenous of the circadian system. The macroscopic effect of such asymmetric interaction is the information that continuously flows from the signal (light state) to the response (evolution of the circadian variables state). As we discussed in previous Chapters 4-5, quantitative definitions of information flow and causal influence are currently under debate[Jam+16; Auc+17], and especially for nonlinear systems we do not have a consensus way of quantifying influences. The idea of this chapter is to quantify the effect of photic perturbations on circadian rhythms considering the modification that they produce in the joint signal-response time series irreversibility. In particular, because of the intrinsic irreversibility of circadian oscillations that is

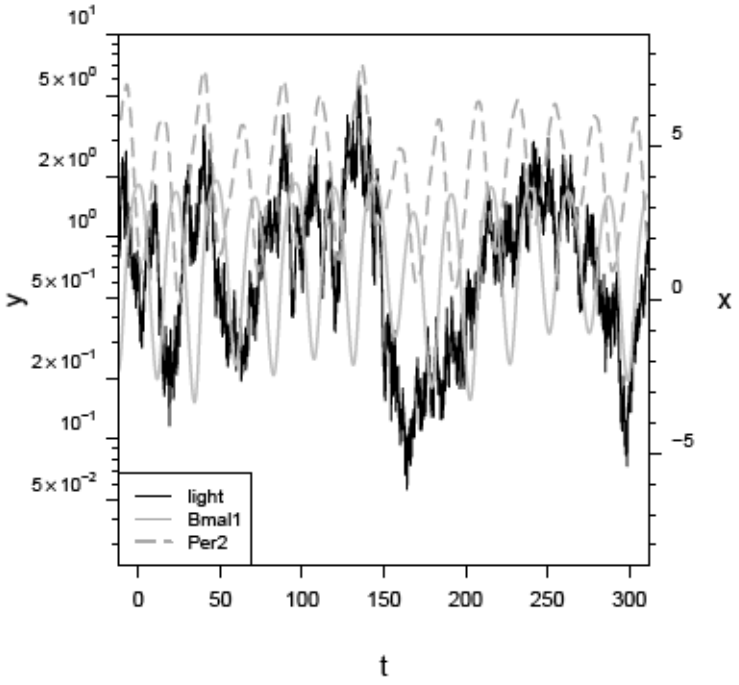


Fig. 6.3: Stochastic dynamics of the circadian model genes $Per2$ and $Bmal1$ perturbed with a light x fluctuations of intensity $\gamma = 0.05$ and relaxation time $t_{rel} = 10h$.

observed in the absence of perturbations, we will use a measure of mutual mapping irreversibility, defined as the Markovian approximation to the mutual entropy production introduced by Diana and Esposito in [DE14].

We model the continuous photic perturbation as multiplicative noise on the production rate of $Per2$ mRNA with intensity parameter γ. We describe the dynamics of light x with correlated fluctuations represented by an Ornstein-Uhlenbeck process[UO30], that is the simplest model of dynamical stationary fluctuations (see Chapter 2). The characteristic time of fluctuations we fix to $t_{rel} = 10h$ (hours), that is compatible with the average time of environmental changes experienced by the circadian clock. The equation system for the light x and genes $\{y_i\}_{i=1,..,5}$ dynamics reads:

$$\begin{cases} dx = -\frac{x}{t_{rel}}dt + dW \\ \frac{dy_i}{dt} = f_i(\{y_j(t-\tau_j)\}_{j=1,...,5}) + \delta_{i2} y_i \gamma x \end{cases} \tag{6.6}$$

where the Kronecker delta δ_{i2} selects the light perturbation to act only on $Per2$. dW represents Brownian motion[Shr04], which is specified by $\langle dW(t_i) dW(t_j) \rangle = \delta_{ij} dt$. The exact form of the regulating functions f_i and the corresponding parameters

values are given in the previous section. A sample realization of the dynamics with light fluctuations intensity $\gamma = 0.05$ is plotted in Fig.6.3.

6.3.1 Spectral analysis

While $Per2$ is directly influenced by light, $Bmal1$ is influenced only indirectly through $Per2 \rightarrow Rev\text{-}erb\alpha \rightarrow Bmal1$ and longer paths. The light perturbation modifies the trajectories from being regular allowing oscillations to occur on different periods than 24 hours. This is seen studying the spectral content of trajectories for different values of the light intensity γ. Let us recall the definition of power spectral density $\mu_y(w)$ of a process y as a function of the frequency w:

$$\mu_y(w) = \lim_{T\to\infty} \frac{\left\langle |\int_0^T dt\ e^{-iwt} y(t)|^2 \right\rangle}{T}. \tag{6.7}$$

We see in Fig.6.4 that the power spectral density of variable $Bmal1$, $\mu_{Bmal1}(w)$, for small values of γ has a sharp peak at around $\frac{1}{24h}$, and that broadens when γ is increased up to values where stable oscillations are practically lost. The effect is even larger on the light sensor $Per2$, this indicating that the light perturbation is propagated into the circadian clock network, and attenuated by the feedback dynamics (6.6) preserving robust oscillations in the other genes[Auc+19a].

6.3.2 The mutual irreversibility quantifies the photic entrainment out of equilibrium

The circadian clock model coupled to the light dynamics (6.6) results to be a stochastic stationary process, at least in the biological parameter ranges. As usual, let us consider the time-invariant joint probability density $p(\zeta_\tau^{xy}) = p(x_t, y_t, x_{t+\tau}, y_{t+\tau})$ of the light x and one of the circadian variables y, taken at two time instants separated by an interval τ. The system is not Markovian due to the time-delayed interactions, therefore the joint probability at two time instants $p(\zeta_\tau^{xy})$ cannot be a complete description of the dynamics. System (6.6) could be expressed in Markovian form if we would consider portions of trajectories, for each variable of a lenght equal to its interaction time delay τ_i. For a time delay of interactions that is comparable to the characteristic time of the dynamics, and this is our case, this approach would not be computationally feasible even for a single variable. If we then consider only the two time points statistics, still the probability density of the 12-dimensional variable $\zeta_\tau^{x\vec{y}}$ cannot be estimated with the precision needed to compare information-theoretic measures. We will therefore consider $p(\zeta_\tau^{xy})$ for one variable y at a time, and varying the observational time τ we wish to gain insights into the light influence propagation through the circadian network. Note that since we consider only one of the circadian

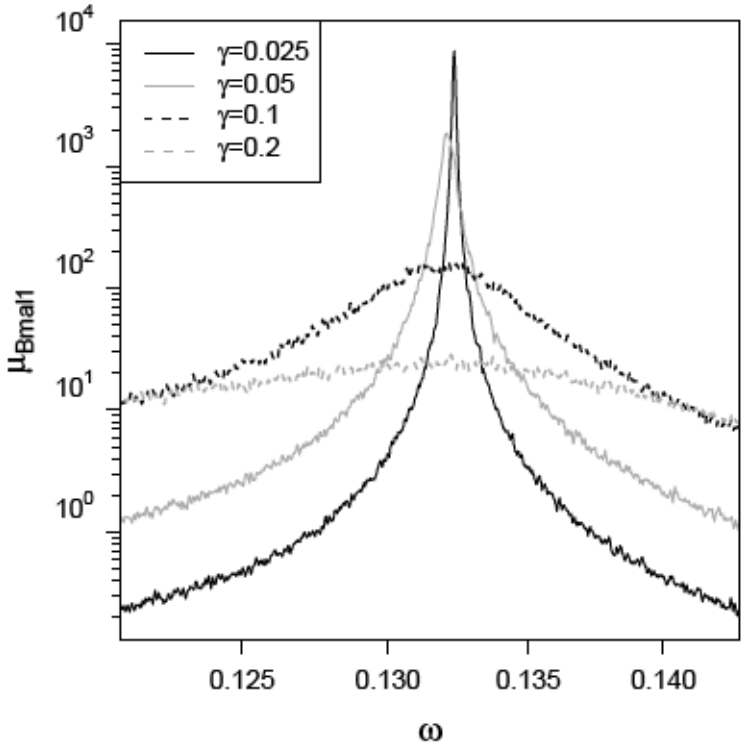


Fig. 6.4: Power spectral density $\mu_{Bmal1}(w)$ of the $Bmal1$ mRNA concentration trajectories, for different values of the light intensity parameter γ.

variables at a time, the conditional probability $p(y_{t+\tau}|x_t, y_t)$ has a larger variance compared to the full knowledge of the state at time t, $p(y_{t+\tau}|x_t, \{y_j(t)\}_{j=1,\dots,5})$.

Recall that, as discussed in Chapter 5, even if the underlying dynamics (6.6) would be bipartite, $p(x_{t+dt}, y_{t+dt}|x_t, y_t) = p(x_{t+dt}|x_t, y_t) \cdot p(y_{t+dt}|x_t, y_t)$, the observation at a finite resolution makes the time series non-bipartite, $p(x_{t+\tau}, y_{t+\tau}|x_t, y_t) = p(x_{t+\tau}|x_t, y_t) \cdot p(y_{t+\tau}|x_t, y_t, x_{t+\tau})$. This is what makes bivariate time series different from the continuous stochastic thermodynamics, in that probabilities cannot be expressed in terms of Onsager-Machlup action functionals[RH16; OM53].

We first explore the effect of perturbations on the shapes of gene oscillations in the dynamics. The irreversibility measured by Φ_τ^y decreases with γ, and this is already an indicator of shape changes. Another aspect could be the distinguishability of the different genes' oscillating trajectories. This requires a measure of distance, like the Kullback-Leibler divergence with probability distributions, but for more general curves this is not an easy task to be formalized precisely. Therefore we took a practical way analyzing the performance of a neural network [Nas07; BD+19] in the problem of recognizing the different genes when it is shown a 3-periods long portion of their normalized dynamics, with a fixed sampling interval $\tau << 24h$. The performance of the neural network is perfect for low γs, and progressively reduces starting from around $\gamma = 0.05$.

Importantly, the circadian oscillations $y(t)$ are time-asymmetric even in the absence of perturbations ($\gamma = 0$) due to the non trivial form of the f_i in (6.6), and this is reflected in the y mapping irreversibility being positive, $\Phi_\tau^y > 0$. The joint irreversibility is lower bounded by that of the subsystems, $\Phi_\tau^{xy} \geq \Phi_\tau^y > 0$, and is therefore not the right measure to quantify the influence of photic perturbations. We wish to neglect the intrinsic asymmetry of such nonlinear oscillations, and to only consider that fraction of irreversibility that results from the continuous photic perturbation. Therefore, in analogy with the definition of mutual entropy production given in [DE14], we define the Markovian approximation to it considering only the statistics of single steps in the time series, and we call it ***mutual mapping irreversibility*** $\Theta_\tau^{xy} \equiv \langle \theta_\tau^{xy} \rangle$. Its stochastic realization-dependent counterpart is written:

$$\theta_\tau^{xy} \equiv \varphi_\tau^{xy} - \varphi_\tau^x - \varphi_\tau^y. \tag{6.8}$$

Θ_τ^{xy} is the amount of mapping irreversibility in the joint time series that is due to the interaction between subsystems. Indeed in the absence of interaction it holds $\theta_\tau^{xy} = 0$.

Our circadian system (6.6) is a signal-response model, because the dynamics of the light x is not affected by any of the circadian variables y_i. As we discussed in Chapter

5, for time series of signal-response models an inequality holds[Auc+19b], posing the backward transfer entropy[Ito16] as a lower bound to the conditional entropy production $\Phi_\tau^{y|x} \equiv \Phi_\tau^{xy} - \Phi_\tau^x \geq T_{y\to x}(-\tau)$. This can be rewritten as a function of the mutual irreversibility as $\Theta_\tau^{xy} + \Phi_\tau^y \geq T_{y\to x}(-\tau)$, but it does not necessarily provide a positive lower bound to the mutual entropy production, since Φ_τ^y is often larger than $T_{y\to x}(-\tau)$. Indeed Θ_τ^{xy} is not defined positive[DE14], and the general lower-bound is $\Theta_\tau^{xy} \geq -\Phi_\tau^{xy}$.

Let us motivate a bit more why we use the mutual irreversibility Θ_τ^{xy} here. We know from previous chapters that in signal-response models the main consequence of the asymmetric interaction between signal and response is the irreversibility of the joint time series. In the BLRM that was clear observing the peaks of the signal often being followed by the peaks of the response, this creating irreversibility in the joint time series. The mutual irreversibility Θ_τ^{xy} is the measure to quantify that kind of effects in the more complex example of circadian oscillations, where nonlinearities and feedbacks govern the dynamics.

Θ_τ^{xy} results to be always positive here, $\Theta_\tau^{xy} \geq 0$. We suspect that this could be a general feature of signal-response models, that the mutual entropy production is non negative. This has to be considered just a speculation though, and we were not able to provide a complete proof. The difficulty comes from the characterization of the anti-causal transfer entropy $\left\langle \ln \frac{p(y_{t+\tau}|x_t,y_t,x_{t+\tau})}{p(y_{t+\tau}|y_t)} \right\rangle$. This problem is somehow related to the one we encountered in optimizing information transmission (see section 3.2), where no simple analytical form was possible for output distributions. Then the small-noise-approximation was introduced taking the output variance just as a propagation of the input variance through the channel. The parallel is not straightforward though since the quantities here are more complicated that the mutual information.

6.3.3 Mutual irreversibility oscillations

In Fig.6.5 we plot Θ_τ^{xy} for the five genes as a function of the observational time τ, for a light fluctuations intensity of $\gamma = 0.05$. Θ_τ^{xy} vanishes for $\tau \to 0$ because of the uncertainty in the dynamics which derives from the other four non considered variables. In particular, for small τ the distribution $p(x_{t+\tau}, y_{t+\tau}|x_t, y_t)$ is bimodal and similar to $p(x_{t-\tau}, y_{t-\tau}|x_t, y_t)$, while $p(x_{t+\tau}, y_{t+\tau}|x_t, \vec{y_t})$ is unimodal and different from $p(x_{t-\tau}, y_{t-\tau}|x_t, \vec{y_t})$.

Θ_τ^{xy} increases for all variables for small τ, much before the delay time of interactions with $Per2$, $\tau < \tau_{Per2}$, because of the correlation time of the signal. In other words, the knowledge of signal state at time t, that is x_t, gives a non-negligible amount

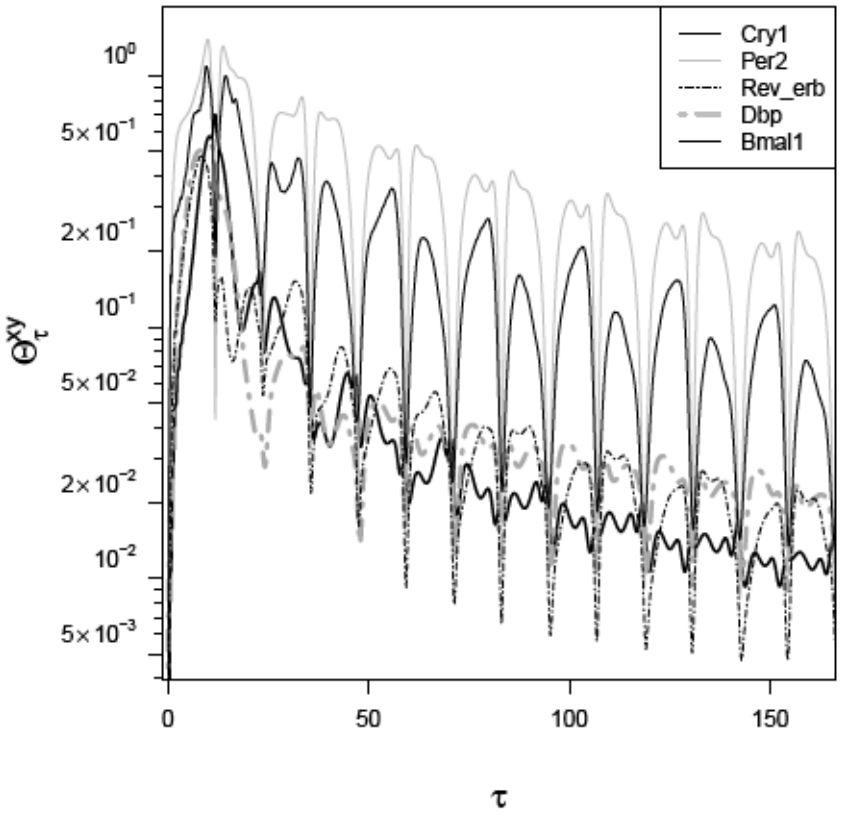


Fig. 6.5: Mutual mapping irreversibility Θ^{xy}_τ for the five circadian variables as a function of the observational time τ, for light fluctuations of intensity $\gamma = 0.05$.

of information on the signal at previous time instants $t - \tau \sim t - t_{rel}$, which then gives information on the other variables at previous times where their delayed influence on time t matters. We see that, after a transient period of roughly 48 hours, the mutual entropy production Θ^{xy}_τ shows periodic regular patterns for all genes while exponentially decaying. We can factor Θ^{xy}_τ assuming the form $\Theta^{xy}_\tau = Ae^{-B\tau} f(\tau)$, where A is the amplitude, B is the decay rate, and $f(\tau)$ is the oscillating component. $Per2$ has the highest intensity $A_{Per2} = 0.58$ being the direct sensor of light fluctuations; it is followed by $Cry1$ with $A_{Cry1} = 0.32$, and this is consistent with $Cry1$ being the only variable that is influenced by all the others. The remaining variables have a much smaller response to light, $A_{Dbp} \approx A_{Rev-erb\alpha} \approx A_{Bmal1} \approx 0.05$. The decay rates are almost equal for all variables $B_{Per2} \approx B_{Cry1} \approx B_{Dbp} \approx B_{Rev-erb\alpha} \approx B_{Bmal1} \approx 0.01$. In order to characterize the oscillating component $f(\tau)$ we study its spectral content with the single realization discrete PSD. The PSD results to have strong peaks for the harmonics corresponding to 12, 24, and 6 hours periods. We extract the characteristic period of the mutual irreversibility oscillations T as a weighted average of the corresponding harmonics,

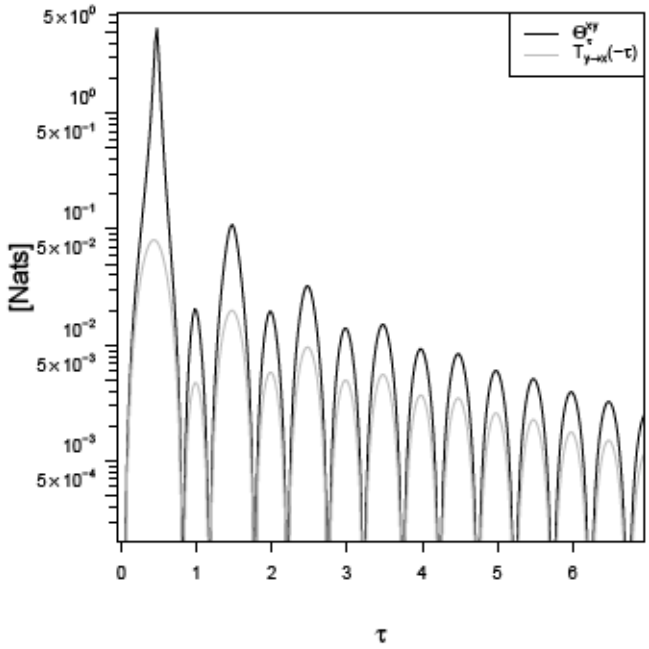


Fig. 6.6: Mutual mapping irreversibility $\Theta_\tau^{x,y}$ in the damped linear oscillator driven by colored noise (6.9), with parameters $t_{rel} = 1$, $\beta = 0.2$, and $\gamma = 1$. In gray we plot the backward transfer entropy $T_{y\to x}(\tau)$, that is the lower bound given by the time series fluctuation theorem [Auc+19b].

$T = \frac{1}{N_{PSD}} \sum_{j=1}^{\infty} \frac{PSD(i)}{w(i)}$ with normalization factor $N_{PSD} = \sum_{j=1}^{\infty} PSD(i)$. The characteristic period of oscillations is around 9-13 hours for all the variables except for $Bmal1$ which has a larger $T_{Bmal1} \approx 33h$ due to its subexponential decay which is reflected in large $PSD(i)$ contributions corresponding to large periods (small i) comparable to the τ axis lenght of 5 days (7 days -2 days of transient behavior). The result $T \approx 12h$ for the other variables means that, while the dynamics is strongly characterized by 24 hours oscillations and 12h harmonics happen only in special cases[WH13], here the response to perturbations induces strong 12 hour harmonics in the mutual irreversibility of all Circadian variables.

The stochastic damped oscillator

Let us show here that the 12 hour harmonics in the mutual irreversibility is not due to the nonlinear behavior, and also not to the self-sustained property. Indeed, they are observed also in a linear damped oscillator y driven by colored noise x:

$$\begin{cases} dx = -\frac{x}{t_{rel}} dt + dW \\ \frac{dy}{dt} = -\beta y + \gamma x \; - (2\pi)^2 \int_{-\infty}^{t} dt' \; y_{t'} e^{-\beta(t-t')} \end{cases} \tag{6.9}$$

where $\beta > 0$ and the term $(2\pi)^2$ sets the oscillations' period to 1. Here the mutual irreversibility peaks occur every half period ($\frac{1}{2}$ units), see Fig.6.6, and that is also the case for mutual information and transfer entropy measures.

In model (6.9) both the subsystem's dynamics is time symmetric, $\Phi_\tau^x = 0$ and $\Phi_\tau^y = 0$, and the irreversibility is seen in the interaction and found in the joint mapping irreversibility Φ_τ^{xy}. Then $\Theta_\tau^{xy} = \Phi_\tau^{xy}$ and the inequality with the backward transfer entropy reads $\Theta_\tau^{xy} \geq T_{y\to x}(\tau)$, and it is plotted in Fig.6.6. Let us also note that, similar to what we found in the circadian clock model (Fig.6.5), the asymmetry of successive peaks decreases with time. The difference in the response to fluctuations between the nonlinear circadian model (6.6) and the linear damped oscillator (6.9) is in the shapes of curves, that look indeed non trivial in the circadian clock mutual irreversibility (Fig.5.10). Another difference is found in the fact that peaks in the linear oscillator mutual irreversibility occur at each half period, while in the circadian genes these correspond to bottom points. The discrepancy could be due to the different nature of oscillations in the dynamics: in the stochastic damped harmonic oscillator these are produced as an integration of the noise source and would not be there in the absence of noise, while in the circadian clock the oscillations are present also in the absence noise where the dynamics approaches a limit cycle. We checked that such bottom points are not a numerical artifact by decreasing the discretization interval in the y axes in the estimation of Θ_τ^{xy}.

7 Conclusions

With this PhD thesis I wished to give a contribution to the study of two fundamental concepts that arise in time series analysis and stochastic modeling: irreversibility and causal influence.

Irreversibility is a measure of time asymmetry. It quantifies how often a process is observed to run backwards, that is to observe particular sequences of events in reversed order. The causal influence is a measure of information flow between objects. It quantifies how much is a variable influencing the dynamics of another variable, and on which time scales the effects of such (asymmetric) interaction are observed.

We adopted a time series framework, because that is the form of data that we get from experiments, where measurements are necessarily performed at a finite frequency. But there is also another reason, the time series framework allows us to analyze the system's dynamics on a time scale of our choice. Indeed the observational time τ is the parameter on which we focused more in all the examples.

Starting from the discussion of bipartite systems, and introducing the main results in the literature, we then set the basis for the information thermodynamics of time series, that are non-bipartite in general, and we found a fluctuation theorem that relates information flow and irreversibility in signal-response models. Importantly, this shows that the connection between information flow and irreversibility in time series is not general, but it can derive from an incomplete causal structure of the dynamics, that is the absence of feedback in signal-response models[Auc+19b]. Particular attention was given to the way in which joint probabilities are factorized in conditional probabilities, and we introduced the causal representation as that Bayesian network describing the (Markovian) time series that minimizes the number of links. We speculate that fluctuation theorems in the time series information thermodynamics arise as a consequence of the incompleteness of such causal representations. A multidimensional generalization of this kind is still missing.

We proposed a new definition of causal influence[Auc+17] as a measure of non-redundant information flow in the partial information decomposition framework. It has a functional form that is imposed by the information processing properties

of linear signal-response models, and by a symmetry requirement. Its behavior in time series of linear signal-response models as a peak function of time vanishing for both limits $\tau \to 0$ and $\tau \to \infty$, reflects the intuition that effects of asymmetric causal interactions are observed gradually over time and then vanish after long enough intervals. We also discussed the three-variables generalization and showed that, if a time-lagged correlation between two variables is only induced by a common input variable and not by direct or indirect interaction, then it is correctly calculated by our measure as zero causal influence. Unfortunately, we are currently unable to extend the definition of causal influence to systems with a general feedback structure and nonlinearities.

A connection between causal influence and irreversibility is found in time series of linear signal-response models for $\tau \to 0$. In that case indeed we showed that the backward transfer entropy converges to the causal influence while being related to the time series irreversibility through the fluctuation theorem.

The particular structure of signal-response models limits a lot the applicability of this study to real world examples, where feedback is often the interesting property. Nevertheless with the circadian clock example we showed that, if those feedback interactions are limited to the internal dynamics of the response, then the theory on signal-response models still allows us to describe the response of feedback systems to perturbations with the language of time series information thermodynamics.

The mutual irreversibility is that part of the joint irreversibility that is due to the interaction between subsystems. We quantified the effect of continuous photic perturbations in a minimal model of the circadian clock network, assuming the mutual irreversibility of time series to be the key effect of such asymmetric interaction[Auc+19a]. The photic perturbation affects the different genes with different intensities depending on their position in the network structure, but all with the same half period harmonics structure. Such harmonics we also found in the response to perturbations of a damped harmonic oscillator, but with an opposite order of peaks and bottoms. Note that contrary to the damped harmonic oscillator, even in the absence of external perturbations the circadian model has a limit cycle, and that is produced by feedback, highly nonlinear, and explicitly delayed interactions between genes.

The physical origin of the multiplicative noise that we used in the circadian clock model and in the receptor-ligand model, has not been well motivated here and will be the object of further studies. Also, the search for efficient methods in the numerical estimation of the time series mutual irreversibility in the circadian clock and other models of such complexity will be the object of further studies.

Bibliography

[Abr+18] Ute Abraham, Julia Katharina Schlichting, Achim Kramer, and Hanspeter Herzel. „Quantitative analysis of circadian single cell oscillations in response to temperature“. In: *PloS one* 13.1 (2018), e0190004 (cit. on p. 120).

[Ahl53] Lars V Ahlfors. „Complex analysis: an introduction to the theory of analytic functions of one complex variable“. In: *New York, London* (1953), p. 177 (cit. on p. 17).

[Ama97] SI Amari. „Information geometry“. In: *Contemporary Mathematics* 203 (1997), pp. 81–96 (cit. on p. 4).

[Auc+17] Andrea Auconi, Andrea Giansanti, and Edda Klipp. „Causal influence in linear Langevin networks without feedback“. In: *Physical Review E* 95.4 (2017), p. 042315 (cit. on pp. 2, 3, 28, 59, 61, 63, 65, 67–70, 72–77, 80, 85, 88, 89, 94, 97, 98, 104, 108, 110, 113, 115, 124, 133).

[Auc+18] Andrea Auconi, Andrea Giansanti, and Edda Klipp. „A fluctuation theorem for time-series of signal-response models with the backward transfer entropy“. In: *arXiv preprint arXiv:1803.05294* (2018) (cit. on pp. 28, 60, 87).

[Auc+19a] Andrea Auconi, Patrick Pett, Edda Klipp, and Hanspeter Herzel. „Influence of photic perturbations on circadian rhythms“. In: *arXiv preprint arXiv:1903.10239* (2019) (cit. on pp. 3, 119, 126, 134).

[Auc+19b] Andrea Auconi, Andrea Giansanti, and Edda Klipp. „Information Thermodynamics for Time Series of Signal-Response Models“. In: *Entropy* 21.2 (2019), p. 177 (cit. on pp. 3, 28, 60, 87, 129, 131, 133).

[Bar+09] Lionel Barnett, Adam B Barrett, and Anil K Seth. „Granger causality and transfer entropy are equivalent for Gaussian variables“. In: *Physical review letters* 103.23 (2009), p. 238701 (cit. on pp. 60, 68, 113).

[Bar15] Adam B Barrett. „Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems“. In: *Physical Review E* 91.5 (2015), p. 052802 (cit. on pp. 2, 5, 28, 61, 68, 69, 85).

[BB86] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*. Vol. 31999. McGraw-Hill New York, 1986 (cit. on p. 13).

[BD+19] Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff. „Learnability can be undecidable“. In: *Nature Machine Intelligence* 1.1 (2019), p. 44 (cit. on p. 128).

[BE10] Christian Van den Broeck and Massimiliano Esposito. „Three faces of the second law. II. Fokker-Planck formulation“. In: *Physical Review E* 82.1 (2010), p. 011144 (cit. on pp. 47, 49).

[Ben+81] Roberto Benzi, Alfonso Sutera, and Angelo Vulpiani. „The mechanism of stochastic resonance“. In: *Journal of Physics A: mathematical and general* 14.11 (1981), p. L453 (cit. on p. 18).

[Ber+14] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. „Quantifying unique information“. In: *Entropy* 16.4 (2014), pp. 2161–2183 (cit. on pp. 61, 85).

[Ber+98] Carlo Bernardini, Orlando Ragnisco, and Paolo Maria Santini. *Metodi matematici della fisica*. Carocci, 1998 (cit. on pp. 14, 17).

[Ber14] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014 (cit. on p. 31).

[Bia17] William Bialek. „Perspectives on theory at the interface of physics and biology“. In: *Reports on Progress in Physics* 81.1 (2017), p. 012601 (cit. on p. 1).

[Bin+05] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, et al. „Transcriptional regulation by the numbers: models“. In: *Current opinion in genetics & development* 15.2 (2005), pp. 116–124 (cit. on p. 122).

[Bla72] Richard Blahut. „Computation of channel capacity and rate-distortion functions“. In: *IEEE transactions on Information Theory* 18.4 (1972), pp. 460–473 (cit. on p. 29).

[Bor+09] Laurence Borgs, Pierre Beukelaers, Renaud Vandenbosch, et al. „Cell “circadian” cycle: new role for mammalian core clock genes“. In: *Cell Cycle* 8.6 (2009), pp. 832–837 (cit. on p. 119).

[Bro+97] Christian Van den Broeck, JMR Parrondo, Raúl Toral, and Ryoichi Kawai. „Nonequilibrium phase transitions induced by multiplicative noise“. In: *Physical Review E* 55.4 (1997), p. 4084 (cit. on p. 19).

[BS05] William Bialek and Sima Setayeshgar. „Physical limits to biochemical signaling“. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.29 (2005), pp. 10040–10045 (cit. on pp. 85, 103).

[BW+04] Sabine Becker-Weimann, Jana Wolf, Hanspeter Herzel, and Achim Kramer. „Modeling feedback loops of the mammalian circadian oscillator“. In: *Biophysical journal* 87.5 (2004), pp. 3023–3034 (cit. on p. 120).

[CD01] Masud Chaichian and A Demichev. „Path integrals in physics. Vol. 1: Stochastic processes and quantum mechanics“. In: (2001) (cit. on pp. 23, 53).

[Cen+00] M Cencini, M Falcioni, E Olbrich, H Kantz, and Angelo Vulpiani. „Chaos or noise: Difficulties of a distinction“. In: *Physical Review E* 62.1 (2000), p. 427 (cit. on p. 121).

[Che+06] Vladimir Y Chernyak, Michael Chertkov, and Christopher Jarzynski. „Path-integral analysis of fluctuation theorems for general Langevin processes“. In: *Journal of Statistical Mechanics: Theory and Experiment* 2006.08 (2006), P08001 (cit. on pp. 24, 49, 87, 91).

[Cil17] S Ciliberto. „Experiments in stochastic thermodynamics: Short history and perspectives“. In: *Physical Review X* 7.2 (2017), p. 021051 (cit. on p. 87).

[Cos+02] Madalena Costa, Ary L Goldberger, and C-K Peng. „Multiscale entropy analysis of complex physiologic time series“. In: *Physical review letters* 89.6 (2002), p. 068102 (cit. on p. 1).

[Cri+18] Andrea Crisanti, Andrea De Martino, and Jonathan Fiorentino. „Statistics of optimal information flow in ensembles of regulatory motifs“. In: *Physical Review E* 97.2 (2018), p. 022407 (cit. on p. 103).

[Cro98] Gavin E. Crooks. „Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems“. In: *Journal of Statistical Physics* 90.5 (1998), pp. 1481–1487 (cit. on pp. 33, 37).

[Cro99] Gavin E Crooks. „Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences“. In: *Physical Review E* 60.3 (1999), p. 2721 (cit. on pp. 4, 87, 90, 107).

[CS19] Gavin E Crooks and Susanne Still. „Marginal and conditional second laws of thermodynamics“. In: *EPL (Europhysics Letters)* 125.4 (2019), p. 40005 (cit. on p. 94).

[CT12] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012 (cit. on pp. 1, 25, 27–29, 65, 66, 80, 88, 91, 95, 109).

[Day+01] Peter Dayan, Laurence F Abbott, and L Abbott. „Theoretical neuroscience: computational and mathematical modeling of neural systems“. In: (2001) (cit. on p. 1).

[DB95] M DeWeese and W Bialek. „Information flow in sensory neurons“. In: *Il Nuovo Cimento D* 17.7-8 (1995), pp. 733–741 (cit. on p. 1).

[DE14] Giovanni Diana and Massimiliano Esposito. „Mutual entropy production in bipartite systems“. In: *Journal of Statistical Mechanics: Theory and Experiment* 2014.4 (2014), P04010 (cit. on pp. 6, 120, 125, 128, 129).

[Din+06] Mingzhou Ding, Yonghong Chen, and Steven L Bressler. „Granger causality: basic theory and application to neuroscience“. In: *Handbook of time series analysis: recent theoretical developments and applications* (2006), pp. 437–460 (cit. on p. 60).

[Dir81] Paul Adrien Maurice Dirac. *The principles of quantum mechanics*. 27. Oxford university press, 1981 (cit. on p. 7).

[Dou+93] John K Douglass, Lon Wilkens, Eleni Pantazelou, and Frank Moss. „Noise enhancement of information transfer in crayfish mechanoreceptors by stochastic resonance“. In: *Nature* 365.6444 (1993), p. 337 (cit. on p. 18).

[DTW12] Stefano Di Talia and Eric F Wieschaus. „Short-term integration of Cdc25 dynamics controls mitotic entry during Drosophila gastrulation“. In: *Developmental cell* 22.4 (2012), pp. 763–774 (cit. on pp. 85, 104).

[DTW14] Stefano Di Talia and Eric F Wieschaus. „Simple biochemical pathways far from steady state can provide switchlike and integrated responses“. In: *Biophysical journal* 107.3 (2014), pp. L1–L4 (cit. on pp. 85, 104).

[Dub+13] Julien O Dubuis, Gašper Tkačik, Eric F Wieschaus, Thomas Gregor, and William Bialek. „Positional information, in bits“. In: *Proceedings of the National Academy of Sciences* (2013), p. 201315642 (cit. on p. 32).

[Duc07] John Duchi. „Derivations for linear algebra and optimization“. In: *Berkeley, California* (2007) (cit. on p. 117).

[Dun+04] Jay C Dunlap, Jennifer J Loros, and Patricia J DeCoursey. *Chronobiology: biological timekeeping*. Sinauer Associates, 2004 (cit. on p. 119).

[EB10] Massimiliano Esposito and Christian Van den Broeck. „Three faces of the second law. I. Master equation formulation“. In: *Physical Review E* 82.1 (2010), p. 011143 (cit. on p. 47).

[Ein56] Albert Einstein. *Investigations on the Theory of the Brownian Movement*. Courier Corporation, 1956 (cit. on p. 8).

[ES02] Denis J Evans and Debra J Searles. „The fluctuation theorem“. In: *Advances in Physics* 51.7 (2002), pp. 1529–1585 (cit. on p. 87).

[FC08] Edward H Feng and Gavin E Crooks. „Length of time's arrow“. In: *Physical review letters* 101.9 (2008), p. 090602 (cit. on p. 87).

[FP03] Daniel B Forger and Charles S Peskin. „A detailed predictive model of the mammalian circadian clock“. In: *Proceedings of the National Academy of Sciences* 100.25 (2003), pp. 14806–14811 (cit. on p. 120).

[Fuh+15] Luise Fuhr, Mónica Abreu, Patrick Pett, and Angela Relógio. „Circadian systems biology: When time matters“. In: *Computational and structural biotechnology journal* 13 (2015), pp. 417–426 (cit. on pp. 119, 120).

[Gam+98] Luca Gammaitoni, Peter Hänggi, Peter Jung, and Fabio Marchesoni. „Stochastic resonance“. In: *Reviews of modern physics* 70.1 (1998), p. 223 (cit. on p. 18).

[Gar09] Crispin Gardiner. *Stochastic methods*. Vol. 4. springer Berlin, 2009 (cit. on pp. 66, 78).

[Ger12] Volker Gerhardt. *Friedrich Nietzsche: Also Sprach Zarathustra*. Vol. 14. Walter de Gruyter, 2012 (cit. on p. vii).

[Gib14] J Willard Gibbs. *Elementary principles in statistical mechanics*. Courier Corporation, 2014 (cit. on p. 8).

[Gil96a] Daniel T Gillespie. „Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral“. In: *Physical review E* 54.2 (1996), p. 2084 (cit. on pp. 11, 63, 98).

[Gil96b] Daniel T Gillespie. „The multivariate langevin and fokker–planck equations“. In: *American Journal of Physics* 64.10 (1996), pp. 1246–1257 (cit. on p. 48).

[Giu+01] Annapaula Giulietti, Lut Overbergh, Dirk Valckx, et al. „An overview of real-time quantitative PCR: applications to quantify cytokine gene expression“. In: *Methods* 25.4 (2001), pp. 386–401 (cit. on p. 1).

[GK14] Virgil Griffith and Christof Koch. „Quantifying synergistic mutual information“. In: *Guided Self-Organization: Inception*. Springer, 2014, pp. 159–190 (cit. on p. 61).

[GM+08] Alex Gomez-Marin, Juan MR Parrondo, and Christian Van den Broeck. „Lower bounds on dissipation upon coarse graining“. In: *Physical Review E* 78.1 (2008), p. 011107 (cit. on p. 93).

[Gol+00] Ary L Goldberger, Luis AN Amaral, Leon Glass, et al. „PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals“. In: *Circulation* 101.23 (2000), e215–e220 (cit. on p. 1).

[GR10] Diego A Golombek and Ruth E Rosenstein. „Physiology of circadian entrainment“. In: *Physiological reviews* 90.3 (2010), pp. 1063–1102 (cit. on pp. 6, 120, 124).

[Gra+09] A Granada, RM Hennig, B Ronacher, A Kramer, and H Herzel. „Phase response curves: elucidating the dynamics of coupled oscillators“. In: *Methods in enzymology* 454 (2009), pp. 1–27 (cit. on pp. 6, 120, 124).

[Gra69] Clive WJ Granger. „Investigating causal relations by econometric models and cross-spectral methods“. In: *Econometrica: Journal of the Econometric Society* (1969), pp. 424–438 (cit. on p. 68).

[Gre+07] Thomas Gregor, David W Tank, Eric F Wieschaus, and William Bialek. „Probing the limits to positional information“. In: *Cell* 130.1 (2007), pp. 153–164 (cit. on p. 32).

[Gri+14] Virgil Griffith, Edwin KP Chong, Ryan G James, Christopher J Ellison, and James P Crutchfield. „Intersection information based on common randomness“. In: *Entropy* 16.4 (2014), pp. 1985–2000 (cit. on pp. 61, 85).

[Ham94] James Douglas Hamilton. *Time series analysis*. Vol. 2. Princeton university press Princeton, NJ, 1994 (cit. on p. 1).

[Har+13] Malte Harder, Christoph Salge, and Daniel Polani. „Bivariate measure of redundant information“. In: *Physical Review E* 87.1 (2013), p. 012130 (cit. on p. 61).

[Har+16] David Hartich, Andre C Barato, and Udo Seifert. „Sensory capacity: An information theoretical measure of the performance of a sensor“. In: *Physical Review E* 93.2 (2016), p. 022116 (cit. on p. 55).

[HE14] Jordan M Horowitz and Massimiliano Esposito. „Thermodynamics with continuous information flow“. In: *Physical Review X* 4.3 (2014), p. 031015 (cit. on pp. 4, 5, 47, 49, 60, 88–91, 98, 113).

[Hor15] Jordan M Horowitz. „Multipartite information flow for multiple Maxwell demons“. In: *Journal of Statistical Mechanics: Theory and Experiment* 2015.3 (2015), P03006 (cit. on p. 88).

[HS14] Jordan M Horowitz and Henrik Sandberg. „Second-law-like inequalities with information and their interpretations“. In: *New Journal of Physics* 16.12 (2014), p. 125007 (cit. on pp. 51, 55, 60, 91, 113, 115).

[Hua87] Kerson Huang. „Statistical Mechanics, 2nd“. In: *Edition (New York: John Wiley & Sons)* (1987) (cit. on pp. 8, 26, 38).

[IS11] Sosuke Ito and Masaki Sano. „Effects of error on fluctuations under feedback control“. In: *Physical Review E* 84.2 (2011), p. 021123 (cit. on p. 45).

[IS13] Sosuke Ito and Takahiro Sagawa. „Information thermodynamics on causal networks“. In: *Physical review letters* 111.18 (2013), p. 180603 (cit. on pp. 2–4, 28, 87, 90, 91, 107).

[IS15] Sosuke Ito and Takahiro Sagawa. „Maxwell's demon in biochemical signal transduction with feedback loop“. In: *Nature communications* 6 (2015), p. 7498 (cit. on pp. 4, 108).

[IS16] Sosuke Ito and Takahiro Sagawa. „Information flow and entropy production on Bayesian networks“. In: *Mathematical Foundations and Applications of Graph Entropy* 3 (2016), pp. 63–100 (cit. on p. 95).

[Ito16] Sosuke Ito. „Backward transfer entropy: Informational measure for detecting hidden Markov models and its interpretations in thermodynamics, gambling and causality“. In: *Scientific reports* 6 (2016) (cit. on pp. 2, 4, 5, 28, 55, 56, 88, 91, 96, 97, 129).

[Ito18] Sosuke Ito. „Unified framework for the second law of thermodynamics and information thermodynamics based on information geometry“. In: *arXiv preprint arXiv:1810.09545* (2018) (cit. on p. 4).

[Jam+16] Ryan G James, Nix Barnett, and James P Crutchfield. „Information Flows? A Critique of Transfer Entropies“. In: *Physical Review Letters* 116.23 (2016), p. 238701 (cit. on pp. 1, 2, 27, 28, 60, 68, 80, 124).

[Jar00] Christopher Jarzynski. „Hamiltonian derivation of a detailed fluctuation theorem“. In: *Journal of Statistical Physics* 98.1-2 (2000), pp. 77–102 (cit. on pp. 87, 107).

[Jar06a] Christopher Jarzynski. „Rare events and the convergence of exponentially averaged work values“. In: *Physical Review E* 73.4 (2006), p. 046105 (cit. on pp. 37–39).

[Jar06b] Christopher Jarzynski. „Rare events and the convergence of exponentially averaged work values“. In: *Phys. Rev. E* 73 (4 2006), p. 046105 (cit. on p. 108).

[Jar11] Christopher Jarzynski. „Equalities and inequalities: irreversibility and the second law of thermodynamics at the nanoscale“. In: *Annu. Rev. Condens. Matter Phys.* 2.1 (2011), pp. 329–351 (cit. on pp. 4, 87).

[Jar97] Christopher Jarzynski. „Nonequilibrium equality for free energy differences“. In: *Physical Review Letters* 78.14 (1997), p. 2690 (cit. on pp. 35, 87).

[Joh99] Carl Hirschie Johnson. „Forty years of PRCs-What have we learned?“ In: *Chronobiology international* 16.6 (1999), pp. 711–743 (cit. on p. 120).

[Kaw+07] R Kawai, JMR Parrondo, and Christian Van den Broeck. „Dissipation: The phase-space perspective“. In: *Physical review letters* 98.8 (2007), p. 080602 (cit. on pp. 3, 28, 37, 87).

[Khi34] Alexander Khintchine. „Korrelationstheorie der stationären stochastischen Prozesse“. In: *Mathematische Annalen* 109.1 (1934), pp. 604–615 (cit. on p. 14).

[Kho06] Boris N Kholodenko. „Cell-signalling dynamics in time and space“. In: *Nature reviews Molecular cell biology* 7.3 (2006), p. 165 (cit. on pp. 85, 103).

[Kir+16] Christoph Kirst, Marc Timme, and Demian Battaglia. „Dynamic information routing in complex networks“. In: *Nature communications* 7 (2016) (cit. on p. 85).

[Kli+16] Edda Klipp, Wolfram Liebermeister, Christoph Wierling, Axel Kowald, and Ralf Herwig. *Systems biology: a textbook*. John Wiley & Sons, 2016 (cit. on pp. 8, 59, 103).

[Koi+12] Nobuya Koike, Seung-Hee Yoo, Hung-Chung Huang, et al. „Transcriptional architecture and chromatin landscape of the core circadian clock in mammals“. In: *Science* 338.6105 (2012), pp. 349–354 (cit. on p. 120).

[Kor+14] Anja Korenčič, Grigory Bordyugov, Robert Lehmann, Damjana Rozman, Hanspeter Herzel, et al. „Timing of circadian genes in mammalian tissues“. In: *Scientific reports* 4 (2014), p. 5782 (cit. on pp. 6, 120–122, 124).

[Kos+14] Jonne V Koski, Ville F Maisi, Jukka P Pekola, and Dmitri V Averin. „Experimental realization of a Szilard engine with a single electron“. In: *Proceedings of the National Academy of Sciences* 111.38 (2014), pp. 13786–13789 (cit. on p. 87).

[Kra+18] Diego Krapf, Enzo Marinari, Ralf Metzler, et al. „Power spectral density of a single Brownian trajectory: what one can and cannot learn from it“. In: *New Journal of Physics* (2018) (cit. on p. 13).

[Kub57] Ryogo Kubo. „Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems“. In: *Journal of the Physical Society of Japan* 12.6 (1957), pp. 570–586 (cit. on p. 67).

[Kub66] Rep Kubo. „The fluctuation-dissipation theorem“. In: *Reports on progress in physics* 29.1 (1966), p. 255 (cit. on p. 67).

[Kur98] Jorge Kurchan. „Fluctuation theorem for stochastic dynamics“. In: *Journal of Physics A: Mathematical and General* 31.16 (1998), p. 3719 (cit. on p. 4).

[Lac+12] Lucas Lacasa, Angel Nunez, Édgar Roldán, Juan MR Parrondo, and Bartolo Luque. „Time series irreversibility: a visibility graph approach“. In: *The European Physical Journal B* 85.6 (2012), p. 217 (cit. on p. 3).

[Lee+15] UnCheol Lee, Stefanie Blain-Moraes, and George A Mashour. „Assessing levels of consciousness with symbolic analysis“. In: *Phil. Trans. R. Soc. A* 373.2034 (2015), p. 20140117 (cit. on p. 28).

[Leh+15] Robert Lehmann, Liam Childs, Philippe Thomas, et al. „Assembly of a comprehensive regulatory network for the mammalian circadian clock: a bioinformatics approach“. In: *PLoS One* 10.5 (2015), e0126283 (cit. on p. 119).

[LG03] Jean-Christophe Leloup and Albert Goldbeter. „Toward a detailed computational model for the mammalian circadian clock“. In: *Proceedings of the National Academy of Sciences* 100.12 (2003), pp. 7051–7056 (cit. on p. 120).

[LG05] Rhonald C Lua and Alexander Y Grosberg. „Practical applicability of the Jarzynski relation in statistical mechanics: A pedagogical example“. In: *The Journal of Physical Chemistry B* 109.14 (2005), pp. 6805–6811 (cit. on pp. 36, 38).

[Liz+08] Joseph T Lizier, Mikhail Prokopenko, and Albert Y Zomaya. „Local information transfer as a spatiotemporal filter for complex systems“. In: *Physical Review E* 77.2 (2008), p. 026110 (cit. on p. 28).

[Lor63] Edward N Lorenz. „Deterministic nonperiodic flow“. In: *Journal of the atmospheric sciences* 20.2 (1963), pp. 130–141 (cit. on p. 7).

[Mar+08] Umberto Marini Bettolo Marconi, Andrea Puglisi, Lamberto Rondoni, and Angelo Vulpiani. „Fluctuation–dissipation: response theory in statistical physics“. In: *Physics reports* 461.4 (2008), pp. 111–195 (cit. on p. 67).

[Mar+09] Koji Maruyama, Franco Nori, and Vlatko Vedral. „Colloquium: The physics of Maxwell's demon and information“. In: *Reviews of Modern Physics* 81.1 (2009), p. 1 (cit. on p. 44).

[Mar+10] Biliana Marcheva, Kathryn Moynihan Ramsey, Ethan D Buhr, et al. „Disruption of the clock components CLOCK and BMAL1 leads to hypoinsulinaemia and diabetes“. In: *Nature* 466.7306 (2010), p. 627 (cit. on p. 119).

[Mar+16] Ignacio A Martínez, Édgar Roldán, Luis Dinis, et al. „Brownian carnot engine“. In: *Nature physics* 12.1 (2016), p. 67 (cit. on p. 87).

[Mas90] James Massey. „Causality, feedback and directed information“. In: *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*. Citeseer. 1990, pp. 303–305 (cit. on pp. 1, 56).

[Maz+12] Gianluigi Mazzoccoli, Valerio Pazienza, and Manlio Vinciguerra. „Clock genes and clock-controlled genes in the regulation of metabolic rhythms“. In: *Chronobiology international* 29.3 (2012), pp. 227–251 (cit. on p. 120).

[McN+88] Bruce McNamara, Kurt Wiesenfeld, and Rajarshi Roy. „Observation of stochastic resonance in a ring laser“. In: *Physical Review Letters* 60.25 (1988), p. 2626 (cit. on p. 18).

[MO53] S Machlup and Lars Onsager. „Fluctuations and irreversible process. II. Systems with kinetic energy“. In: *Physical Review* 91.6 (1953), p. 1512 (cit. on pp. 24, 53).

[MR12] T Munakata and ML Rosinberg. „Entropy production and fluctuation theorems under feedback control: the molecular refrigerator model revisited“. In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.05 (2012), P05010 (cit. on pp. 51, 55).

[MR13] T Munakata and ML Rosinberg. „Feedback cooling, measurement errors, and entropy production“. In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.06 (2013), P06014 (cit. on pp. 51, 55).

[MR14] T Munakata and ML Rosinberg. „Entropy production and fluctuation theorems for Langevin processes under continuous non-Markovian feedback control“. In: *Physical review letters* 112.18 (2014), p. 180601 (cit. on p. 55).

[MS87] C. Mencuccini and V. Silvestrini. *Fisica 1. Meccanica termodinamica. Corso di fisica per le facoltà scientifiche. Con esempi ed esercizi*. Argomenti di fisica. Liguori, 1987 (cit. on p. 39).

[MW14] W Moon and JS Wettlaufer. „On the interpretation of Stratonovich calculus“. In: *New Journal of Physics* 16.5 (2014), p. 055017 (cit. on p. 24).

[Nas07] Nasser M Nasrabadi. „Pattern recognition and machine learning“. In: *Journal of electronic imaging* 16.4 (2007), p. 049901 (cit. on p. 128).

[Nem12] Ilya Nemenman. „Gain control in molecular information processing: lessons from neuroscience“. In: *Physical biology* 9.2 (2012), p. 026003 (cit. on pp. 65, 104).

[Nie98] Friedrich Wilhelm Nietzsche. *Thus spoke zarathustra*. Prabhat Prakashan, 1898 (cit. on p. vii).

[Oga88] Yosihiko Ogata. „Statistical models for earthquake occurrences and residual analysis for point processes“. In: *Journal of the American Statistical association* 83.401 (1988), pp. 9–27 (cit. on p. 1).

[OM53] Lars Onsager and S Machlup. „Fluctuations and irreversible processes“. In: *Physical Review* 91.6 (1953), p. 1505 (cit. on pp. 24, 53, 128).

[Ovc16] Igor V Ovchinnikov. „Introduction to supersymmetric theory of stochastics“. In: *Entropy* 18.4 (2016), p. 108 (cit. on p. 24).

[Par+09] Juan MR Parrondo, Christian Van den Broeck, and Ryoichi Kawai. „Entropy production and the arrow of time“. In: *New Journal of Physics* 11.7 (2009), p. 073008 (cit. on pp. 37, 87).

[Par+15] Juan MR Parrondo, Jordan M Horowitz, and Takahiro Sagawa. „Thermodynamics of information“. In: *Nature physics* 11.2 (2015), pp. 131–139 (cit. on pp. 2, 28, 60, 91).

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009 (cit. on p. 62).

[Pea95] Judea Pearl. „Causal diagrams for empirical research“. In: *Biometrika* 82.4 (1995), pp. 669–688 (cit. on p. 62).

[Pet+16] J Patrick Pett, Anja Korenčič, Felix Wesener, Achim Kramer, and Hanspeter Herzel. „Feedback loops of the mammalian circadian clock constitute repressilator“. In: *PLoS computational biology* 12.12 (2016), e1005266 (cit. on p. 124).

[Phi+12] Rob Phillips, Julie Theriot, Jane Kondev, and Hernan Garcia. *Physical biology of the cell*. Garland Science, 2012 (cit. on p. 51).

[Por+07] A Porporato, JR Rigby, and E Daly. „Irreversibility and fluctuation theorem in stationary time series“. In: *Physical review letters* 98.9 (2007), p. 094101 (cit. on p. 3).

[Rau+14] Johannes Rauh, Nils Bertschinger, Eckehard Olbrich, and Jurgen Jost. „Reconsidering unique information: Towards a multivariate information decomposition“. In: *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE. 2014, pp. 2232–2236 (cit. on pp. 2, 68, 76, 85).

[Rei+18] Matthias Reis, Justus A Kromer, and Edda Klipp. „General solution of the chemical master equation and modality of marginal distributions for hierarchic first-order reaction networks“. In: *Journal of mathematical biology* (2018), pp. 1–43 (cit. on p. 29).

[Rel+11] Angela Relógio, Pal O Westermark, Thomas Wallach, et al. „Tuning the mammalian circadian clock: robust synergy of two loops“. In: *PLoS computational biology* 7.12 (2011), e1002309 (cit. on p. 120).

[RH16] Martin Luc Rosinberg and Jordan M Horowitz. „Continuous information flow fluctuations“. In: *EPL (Europhysics Letters)* 116.1 (2016), p. 10007 (cit. on pp. 2, 51–53, 88, 90, 91, 115, 128).

[Ris84] Hannes Risken. „Fokker-planck equation“. In: *The Fokker-Planck Equation*. Springer, 1984, pp. 63–95 (cit. on p. 48).

[Ris96] Hannes Risken. „Fokker-planck equation“. In: *The Fokker-Planck Equation*. Springer, 1996, pp. 63–95 (cit. on p. 90).

[RK17] Jesper C Romers and Marcus Krantz. „rxncon 2.0: a language for executable molecular systems biology“. In: *bioRxiv* (2017), p. 107136 (cit. on p. 121).

[Ros+16] Martin Luc Rosinberg, Gilles Tarjus, and Toyonori Munakata. „Heat fluctuations for underdamped Langevin dynamics“. In: *EPL (Europhysics Letters)* 113.1 (2016), p. 10007 (cit. on p. 54).

[RP12] Édgar Roldán and Juan MR Parrondo. „Entropy production and Kullback-Leibler divergence between stationary trajectories of discrete systems“. In: *Physical Review E* 85.3 (2012), p. 031129 (cit. on pp. 3, 5, 88, 91, 93, 97, 108).

[RW01] Steven M Reppert and David R Weaver. „Molecular analysis of mammalian circadian rhythms“. In: *Annual review of physiology* 63.1 (2001), pp. 647–676 (cit. on pp. 6, 124).

[RW99] Fred Rieke and David Warland. *Spikes: exploring the neural code*. MIT press, 1999 (cit. on p. 14).

[Sag11] Takahiro Sagawa. „Hamiltonian derivations of the generalized Jarzynski equalities under feedback control“. In: *Journal of Physics: Conference Series*. Vol. 297. 1. IOP Publishing. 2011, p. 012015 (cit. on p. 2).

[San+82] José M Sancho, M San Miguel, SL Katz, and JD Gunton. „Analytical and numerical studies of multiplicative noise“. In: *Physical Review A* 26.3 (1982), p. 1589 (cit. on p. 19).

[Sch+03] Elad Schneidman, William Bialek, and Michael J Berry. „Synergy, redundancy, and independence in population codes“. In: *the Journal of Neuroscience* 23.37 (2003), pp. 11539–11553 (cit. on pp. 74, 85).

[Sch00] Thomas Schreiber. „Measuring information transfer“. In: *Physical review letters* 85.2 (2000), p. 461 (cit. on pp. 1, 27, 60, 67, 83).

[Sch12] Arthur Schopenhauer. *The world as will and representation*. Vol. 1. Courier Corporation, 2012 (cit. on pp. 60, 62).

[Sei05] Udo Seifert. „Entropy production along a stochastic trajectory and an integral fluctuation theorem“. In: *Physical review letters* 95.4 (2005), p. 040602 (cit. on pp. 51, 52, 90, 107).

[Sei12] Udo Seifert. „Stochastic thermodynamics, fluctuation theorems and molecular machines“. In: *Reports on Progress in Physics* 75.12 (2012), p. 126001 (cit. on pp. 2, 10, 36, 49, 90, 94, 104).

[Sek10] Ken Sekimoto. *Stochastic energetics*. Vol. 799. Springer, 2010 (cit. on p. 53).

[Sek98] Ken Sekimoto. „Langevin equation and thermodynamics“. In: *Progress of Theoretical Physics Supplement* 130 (1998), pp. 17–27 (cit. on pp. 53, 88, 90, 91).

[Sha01] Claude Elwood Shannon. „A mathematical theory of communication“. In: *ACM SIGMOBILE mobile computing and communications review* 5.1 (2001), pp. 3–55 (cit. on p. 25).

[Shr04] Steven E Shreve. *Stochastic calculus for finance II: Continuous-time models*. Vol. 11. Springer Science & Business Media, 2004 (cit. on pp. 8, 62, 90, 96, 98, 104, 125).

[Shr12] Steven Shreve. *Stochastic calculus for finance I: the binomial asset pricing model*. Springer Science & Business Media, 2012 (cit. on pp. 8, 21).

[Spi+16] Richard E Spinney, Joseph T Lizier, and Mikhail Prokopenko. „Transfer entropy in physical systems and the arrow of time". In: *Physical Review E* 94.2 (2016), p. 022135 (cit. on pp. 88, 91).

[SSC09] Saurabh Sahar and Paolo Sassone-Corsi. „Metabolism and cancer: the circadian clock connection". In: *Nature Reviews Cancer* 9.12 (2009), p. 886 (cit. on p. 119).

[ST15] Thomas R Sokolowski and Gašper Tkačik. „Optimizing information flow in small genetic networks. IV. Spatial coupling". In: *Physical Review E* 91.6 (2015), p. 062710 (cit. on p. 1).

[Sti+12] Susanne Still, David A Sivak, Anthony J Bell, and Gavin E Crooks. „Thermodynamics of prediction". In: *Physical review letters* 109.12 (2012), p. 120604 (cit. on p. 3).

[Str+98] Steven P Strong, Roland Koberle, Rob R de Ruyter van Steveninck, and William Bialek. „Entropy and information in neural spike trains". In: *Physical review letters* 80.1 (1998), p. 197 (cit. on p. 1).

[SU09] Takahiro Sagawa and Masahito Ueda. „Minimal energy cost for thermodynamic information processing: Measurement and information erasure". In: *Physical review letters* 102.25 (2009), p. 250602 (cit. on p. 4).

[SU10] Takahiro Sagawa and Masahito Ueda. „Generalized Jarzynski equality under nonequilibrium feedback control". In: *Physical review letters* 104.9 (2010), p. 090602 (cit. on pp. 40, 87, 90).

[SU12] Takahiro Sagawa and Masahito Ueda. „Nonequilibrium thermodynamics of feedback control". In: *Physical Review E* 85.2 (2012), p. 021104 (cit. on pp. 2, 4, 40, 42–44, 56, 87, 90).

[SU13] Takahiro Sagawa and Masahito Ueda. „Information Thermodynamics: Maxwell's Demon in Nonequilibrium Dynamics". In: *Nonequilibrium Statistical Physics of Small Systems: Fluctuation Relations and Beyond* (2013), pp. 181–211 (cit. on pp. 39, 40).

[Szi64] Leo Szilard. „On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings". In: *Systems Research and Behavioral Science* 9.4 (1964), pp. 301–310 (cit. on pp. 40, 87).

[Tay08] Stephen J Taylor. *Modelling financial time series*. world scientific, 2008 (cit. on p. 1).

[TC07] Tooru Taniguchi and EGD Cohen. „Onsager-Machlup theory for nonequilibrium steady states and fluctuation theorems". In: *Journal of Statistical Physics* 126.1 (2007), pp. 1–41 (cit. on p. 90).

[Thi+17] Sebastian Thieme, Jesper C Romers, Ulrike Muenzner, and Marcus Krantz. „Bipartite Boolean modelling-a method for mechanistic simulation and validation of large-scale signal transduction networks“. In: *bioRxiv* (2017), p. 107235 (cit. on p. 121).

[Tka+08a] Gašper Tkačik, Curtis G Callan Jr, and William Bialek. „Information capacity of genetic regulatory elements“. In: *Physical Review E* 78.1 (2008), p. 011910 (cit. on p. 29).

[Tka+08b] Gašper Tkačik, Curtis G Callan, and William Bialek. „Information flow and optimization in transcriptional regulation“. In: *Proceedings of the National Academy of Sciences* 105.34 (2008), pp. 12265–12270 (cit. on pp. 4, 8, 29, 30, 32, 103).

[Tka+09] Gašper Tkačik, Aleksandra M Walczak, and William Bialek. „Optimizing information flow in small genetic networks“. In: *Physical Review E* 80.3 (2009), p. 031920 (cit. on pp. 1, 8, 104).

[Tka+12a] Gašper Tkacik, Aleksandra M Walczak, and William Bialek. „Optimizing information flow in small genetic networks. III. A self-interacting gene“. In: *Phys. Rev. E* 85.4 (2012), p. 041903 (cit. on p. 1).

[Tka+12b] Gašper Tkačik, Aleksandra M Walczak, and William Bialek. „Optimizing information flow in small genetic networks. III. A self-interacting gene“. In: *Physical Review E* 85.4 (2012), p. 041903 (cit. on p. 8).

[Toy+10] Shoichi Toyabe, Takahiro Sagawa, Masahito Ueda, Eiro Muneyuki, and Masaki Sano. „Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality“. In: *Nature Physics* 6.12 (2010), pp. 988–992 (cit. on pp. 40, 87).

[UO30] George E Uhlenbeck and Leonard S Ornstein. „On the theory of the Brownian motion“. In: *Physical review* 36.5 (1930), p. 823 (cit. on pp. 11, 63, 98, 125).

[Van+06] Katja Vanselow, Jens T Vanselow, Pål O Westermark, et al. „Differential effects of PER2 phosphorylation: molecular basis for the human familial advanced sleep phase syndrome (FASPS)“. In: *Genes & development* 20.19 (2006), pp. 000–000 (cit. on p. 119).

[VK92] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*. Vol. 1. Elsevier, 1992 (cit. on pp. 14, 66).

[Vul10] Angelo Vulpiani. *Chaos: from simple models to complex systems*. Vol. 17. World Scientific, 2010 (cit. on pp. 7, 121).

[Wal+10] Aleksandra M Walczak, Gašper Tkačik, and William Bialek. „Optimizing information flow in small genetic networks. II. Feed-forward interactions“. In: *Physical Review E* 81.4 (2010), p. 041905 (cit. on pp. 1, 8).

[WB10] Paul L Williams and Randall D Beer. „Nonnegative decomposition of multivariate information“. In: *arXiv preprint arXiv:1004.2515* (2010) (cit. on pp. 2, 28, 61, 68, 76, 80).

[Wes+09] Pål O Westermark, David K Welsh, Hitoshi Okamura, and Hanspeter Herzel. „Quantification of circadian rhythms in single cells“. In: *PLoS computational biology* 5.11 (2009), e1000580 (cit. on p. 16).

[WH13] Pål O Westermark and Hanspeter Herzel. „Mechanism for 12 hr rhythm generation by the circadian clock“. In: *Cell reports* 3.4 (2013), pp. 1228–1238 (cit. on pp. 122, 131).

[Wie30] Norbert Wiener. „Generalized harmonic analysis“. In: *Acta mathematica* 55.1 (1930), pp. 117–258 (cit. on p. 14).

[WK11] Christian Waltermann and Edda Klipp. „Information theory based approaches to cellular signaling“. In: *Biochimica et Biophysica Acta (BBA)-General Subjects* 1810.10 (2011), pp. 924–932 (cit. on p. 103).

[Yan09] Lily Yan. „Expression of clock genes in the suprachiasmatic nucleus: effect of environmental lighting conditions“. In: *Reviews in endocrine and metabolic disorders* 10.4 (2009), pp. 301–310 (cit. on pp. 6, 124).

[Zha+14] Ray Zhang, Nicholas F Lahens, Heather I Ballance, Michael E Hughes, and John B Hogenesch. „A circadian gene expression atlas in mammals: implications for biology and medicine“. In: *Proceedings of the National Academy of Sciences* 111.45 (2014), pp. 16219–16224 (cit. on p. 120).

[R C14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2014 (cit. on p. 99).

List of Figures

Acknowledgements

First of all, I would like to thank my supervisor Edda Klipp. She supported me for three years, and she would let me explore the field of information thermodynamics in a independent but rigorous way.

Then I would like to thank my co-supervisor in Rome, Andrea Giansanti, that helped to put my work in perspective and was always available for meaningful discussions.

Importantly I need to say thanks to all my office collaborators, for scientific and non-scientific time spent together, and particularly to Wolfgang and Matthias.

Then I would like to thank the computing and algorithms expert Marco Scazzocchio, for helping with the design of numerical experiments.

Also I need to thank my group of italian friends in Berlin, and especially Valentina, Daniele, Francesco and Gaetano, Viky, Adele and many more.

Special thanks goes to my sister Marina, and also to my best friends in Rome, especially Simone, Luca, Diego, Marco, Michele, Iacopo and Federico.

Declaration

I hereby certify that this thesis has been composed by me and is based on my own work, unless stated otherwise. No other person's work has been used without due acknowledgment in this thesis. All references have been quoted and all sources of information, including graphs and data sets, have been specifically acknowledged.

Berlin, 08.11.2018

Andrea Auconi